Analysis of RNA Energy Folding Landscapes

Day, Luke John

Awarding institution:
King's College London

Analysis of RNA Energy Folding Landscapes

Luke Day

Department of Informatics
King's College London

Submitted in partial fulfillment for
the requirements for the degree of
Doctor of Philosophy

September 2017

Abstract

Over recent years there has been explosive growth in genomic sequence data largely due to rapid advances in sequencing techniques and decreasing costs. It is widely assumed that decoding genomic sequence data will lead to significant advances in our understanding of disease pathogenesis and thus open up new possibilities to tackle diseases. The functional importance of ribonucleic acids (RNAs) as regulatory molecules in normal and abnormal biology has grown considerably. In comparison to proteins, one area of research where there has been slower progress is experimental determination of RNA structure. Based on the fundamental idea that sequence determines a molecule's structure which in turn provides important insights into its biological functions, knowledge of RNA structure is growing in importance.

In this thesis, RNA secondary structures and their folding landscapes are analysed by means of computational techniques. Firstly, we analyse the accessibility of microRNA binding sites over metastable conformations in the context of single nucleotide polymorphisms (SNPs). We developed a tool, `MSbind`, to analyse features of metastable SNP/miRNA binding sites and discovered three parameters that distinguish between alleles. We then incorporated our findings into a new miRNA target site prediction tool, `RNAStrucTar`, which takes into consideration metastable target site accessibility and found, for 16 of 20 [mRNA/3' UTR; SNP; miRNA] instances, `RNAStrucTar` supports experimental findings.

Secondly, we compared random and deterministic descent strategies in the context of RNA folding landscapes. We analyse the speed-up achievable by randomised descent in attraction basins in the context of large sample sets. For the two nongradient methods we analysed for partial energy landscapes induced by ten different RNA sequences; we obtained that the number of observed local minima is on average larger by 7.3% and 3.5%, respectively. The run-time improvement is approximately 16.6% and 6.8% on average over the ten partial energy landscapes. Finally, we propose a new heuristic method based on the general framework devised by Garnier and Kallel to approximate the number of local minima states within partial RNA energy folding landscapes. Our heuristic method achieves for best approximations on average a deviation below 3.0% from the true number of local minima.

Acknowledgements

I owe many thanks to my supervisors Dr Kathleen Steinhöfel and Prof. Tomasz Radzik for their help and guidance over these past few years. I would also like to express my gratitude and appreciation to Prof. Andreas Albrecht for sharing his expertise and insight on computational RNA biology and for collaborating and advising me thoughout this research. I would also like to thank King's College London and the EPSRC for funding this work. Lastly, I would like to thank and express my appreciation to my friends and family who have provided great support over these past few years. Especially, to my parents who have been a constant source of encouragement and support throughout my life.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

The human genome, stored by *deoxyribonucleic acid* (DNA), consists of approximately 3 billion base pairs that encode all 'instructions' required for life. Since completion of the *Human Genome Project* (HGP) back in 2003 huge amounts of genomic sequencing data has become available. And, as genome sequencing technologies continue to improve and costs decrease this data is set to grow at a phenomenal rate in the years ahead. Analyses of this genomic data is expected to elucidate why a unique individual develops a specific disease; providing new insights into questions such as why does cancer develop in manifold ways or why is it the case that only some individuals develop Alzheimer's disease? With the advent of massively parallel sequencing technologies genomic-based diagnostics is rapidly being implemented in clinical practice. For example, *Genomics England* set up in 2013 the 100k genome project which in collaboration with *National Health Service (NHS) England* will sequence the genomes of approximately 75,000 patients by 2017 [1]. With a project mission to develop understanding of disease by bridging the gap between scientific discovery, clinical diagnosis and personalised medicine in a clinical setting. When the first human genome was sequenced the project cost approximately $2.7 billion and took approximately 13 years to complete [2]. During the period of 2001 and 2007 sequencing costs closely followed a Moore's law trajectory, which describes advancement of integrated circuit development. Since the introduction and advancement of next-generation sequencing (NGS) technologies starting in 2008 the costs to sequence genomes have departed from a Moore's law trajectory. For example, the cost to sequence

a genome has dropped from approximately $7 million in 2007 to $1,000 in 2015, see figure 1.1. And, typically requires just over 24 hours thus allowing for larger population-scale sequencing projects.

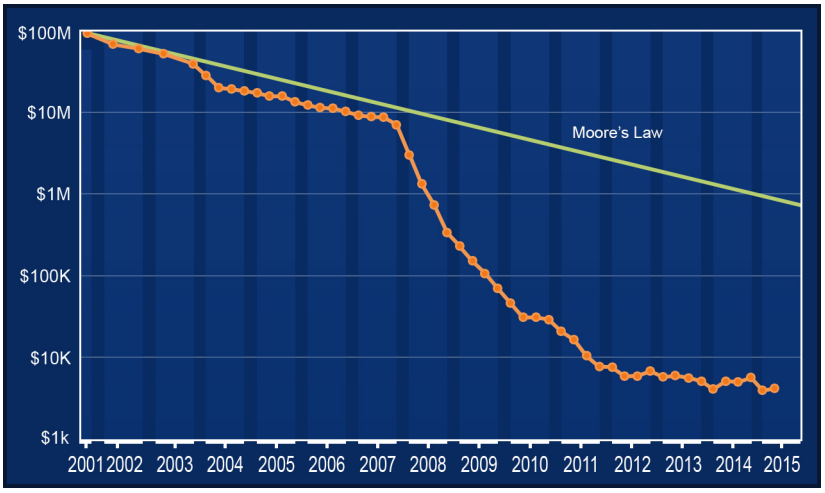

Fig. 1.1 Estimate cost to sequence a human (size = 3,000 Mb) genome. Data from National Human Genome Research Institute (genome.gov/sequencingcosts). 2001 - 2007: *Sanger*-based first generation sequencing platforms. 2008 onwards next generation sequencing platforms.

A major open challenge is decoding the functional complexity of genomic data. Analyses of the significantly growing genomic data is crucial to understanding disease and for the development of future medicines. A central tenet of biology is that structure determines function and conversely knowledge of function typically provides insight into structure. Therefore, one important way genomic sequence data needs to be analysed is at the structural level because, ultimately, it is structure that determines a molecule's function. After discovery of the DNA double helix in 1953 by James Watson and Francis Crick, the primary focus of structural biology research had been on predicting the structures and functions of proteins. Shortly after completion of the HGP the *Encyclopedia of DNA Elements* (ENCODE) consortium began to analyse the genomic data to determine functional elements. Their findings resulted in a major paradigm shift when it was discovered that only around 2% of the human genome consists of protein-coding elements [3].

The precise number of human protein-coding elements has been subject to constant revision ever since discovery of the DNA double helix. Prior to the start of the HGP many researchers estimated the number of protein-coding genes to be in the region of hundreds of thousands. When the initial human genome was published in 2001 the *International Human Genome Sequencing*

Consortium estimated the number to be between 26,000 and 30,000. And, in 2004 when the final draft was published this range was reduced even further to be between 20,000 and 25,000. The latest estimate by GENCODE is just 19,950 human protein-coding genes, representing an extremely small portion of the total human genome (Ensembl 86, version 25, March 2016) [4]. This result is a significant finding because it means the number of protein-coding genes alone do not determine the complexity of an organism. For example, the genome of the sexually transmitted infection *Trichomonas Vaginalis* (T. Vaginalis), a single-cellular parasite, has approximately 60,000 protein-coding genes [5].

In 2012 the ENCODE project released a series of articles in which they approximated that at around 75% of our genome is pervasively transcribed into *ribonucleic acid* (RNA), which is of largely unknown function [6]. A major open question is if 75% of our genome is pervasively transcribed but only 2% codes for protein what is the functional significance of all this RNA? This question is hugely controversial and subject of much debate by biologists. On one side are those who believe the majority of this transcribed RNA to be evolutionary junk or noise [7] and on the other side are those who believe this RNA to contribute to critical, yet to be discovered, functions [8]. What is evident is that for many decades the spotlight has been fixated on proteins but since advancement of NGS technologies this spotlight is increasingly being placed on the importance of the functional and regulatory roles played by RNAs in biological complexity.

1.2 Contributions & thesis structure

Understanding the structures of molecules is crucial to decoding their sequence, structure and function relationship. However, determining structure is particularly difficult for RNA molecules in comparison to say protein molecules. One reason for this is experimental evidence shows RNAs do not always fold into a single, static structure. In this thesis, our specific interests on the analyses of metastable RNA secondary structures, their energy folding landscapes and RNA-RNA interactions is presented. The remaining contents of this thesis is structured as follows:

In Chapter 2 essential background on the chemistry and function of RNA molecules and how this chemistry determines structure is presented. Here we define the difference between coding and non-coding RNAs and describe how RNAs can regulate gene expression. Crucially, the three levels of RNA structure are defined and methods to represent structure are presented.

In Chapter 3 a comprehensive literature overview of computational RNA secondary structure prediction algorithms, their energy folding landscapes, and interaction with other RNAs is discussed. We begin this chapter by exploring the general computational frameworks to predict a single secondary structure; with a focus on the algorithmic ideas underlying the most popular energy minimization framework. We then consider work on RNA folding over energy landscapes and on the computation of metastable conformations. Lastly we review work on predicting the interaction of two RNA molecules with a focus on mRNA-miRNA target site prediction. The aims of this chapter are: (1) to introduce algorithms and tools commonly applied to predict optimal and suboptimal RNA secondary structures, (2) introduce algorithms for predicting the interaction of two RNA molecules with a focus on mRNA-miRNA, and (3) highlight limitations of the computational approaches to (a) predict structure and (b) interaction of RNAs.

Chapter 4 firstly discusses literature on the impact of single nucleotide polymorphisms (SNPs) on gene expression and disease susceptability. Then, I present a new tool `MSbind` used to analyse the impact of SNPs on microRNA target site accessibility over metastable conformations taking into consideration messenger RNA concentration levels. In this work we compiled a dataset from published literature studies on differentiation in gene expression defined by a SNP associated with increased risk to develop specific diseases. Using the dataset and `MSbind` we investigated if the studies reporting increased or decreased expression for the SNP-variant gene have more accessible binding sites in comparison to the non-SNP gene. We then discuss how incorporating metastable structures into miRNA target site prediction can improve prediction accuracy. To conclude this chapter I discuss the importance of metastable secondary structure accessibility on computational microRNA target site prediction. The contents of this chapter appear in the following publication:

[9] **Luke Day, Ouala Abdelhadi Ep Souki, Andreas A. Albrecht and Kathleen Steinhöfel**
Accessibility of microRNA binding sites in metastable RNA secondary structures in the presence of SNPs. Bioinformatics, 30 (3): 343-352, 2014. doi: 10.1093/bioinformatics/btt695.

Chapter 5 Based on our findings presented in the previous chapter we present a new microRNA target site prediction tool, `RNAStrucTar` that takes into consideration metastable RNA secondary structures. We analyse the prediction tool in the context of single nucleotide polymorphisms and compare it to existing methods. The contents of this chapter appear in the following publications:

[10] **Ouala Abdelhadi Ep Souki, Luke Day, Andreas A. Albrecht and Kathleen Steinhöfel**
MicroRNA Target Prediction Based Upon Metastable RNA Secondary Structures. Bioinformatics and Biomedical Engineering, vol 9044 of LNCS, pages 456 - 467, Springer, 2015. doi: 10.1007/978-3-319-16480-$9_4$5

Chapter 6 we compare random and determinstic descent strategies over RNA energy folding landscapes. In this chapter we focus on the comparison of run-time and local minima coverage achieved by the descent strategies. The contents of this chapter appear in the following publication:

[11] **Luke Day, Ouala Abdelhadi Ep Souki, Andreas A. Albrecht and Kathleen Steinhöfel**
Random versus deterministic descent in RNA energy landscape analysis. Article ID 9654921, Advances in Bioinformatics, 2016. doi: 10.1155/2016/9654921

Chapter 7 introduces a new heuristic to approximate the number of local minima in partial RNA energy folding landscapes. Here we present and evaluate the approximation procedure using the three descent strategies discussed in Chapter 6. We conclude this chapter by discussion of potential applications of the approximation heuristic. The contents of this chapter appear in the following publication:

[12] **Andreas A. Albrecht, Luke Day, Ouala Abdelhadi Ep Souki and Kathleen Steinhöfel**
A new heuristic method for approximating the number of local minima in

partial RNA energy landscapes. Computational Biology and Chemistry, 60, pages 43 - 52, 2016. doi: 10.1016/j.compbiolchem.2015.11.002

In Chapter 8 we conclude our work by reviewing its contributions and discussion of the future of computational RNA structure prediction.

Chapter 2

RNA chemistry, structure and function

This chapter introduces the basic chemistry and function of RNA molecules and defines RNA structure.

2.1 Chemistry of RNA molecules

Nucleic acids, *deoxyribonucleic acids* (DNA) and *ribonucleic acids* (RNA), are composed from chemical units called nucleotides (nt.). A nucleotide consists of a pentose or 5-carbon sugar, a phosphate backbone and a nitrogenous base. There are four nitogenous bases found in RNA: *adenine* (A), *guanine* (G), *cytosine* (C) and *uracil* (U). In DNA, the base uracil (U) is replaced with thymine (T).

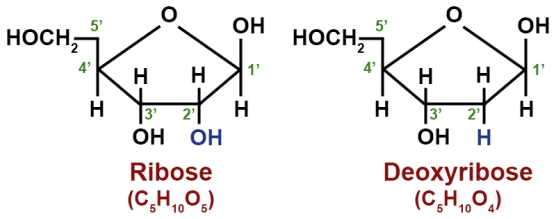

Fig. 2.1 *Ribose sugars*

Chemically the pentose sugars of DNA and RNA are very similar, they both consist of 5 carbon and 10 hydrogen atoms. However in the case of DNA, as its name deoxy suggests it has one less oxygen atom. This difference has a major impact on the flexibility and stability of RNA and as a result difference

in function between DNA and RNA. The 2' hydroxyl is important for allowing RNAs to interact with other molecules. Furthermore, experimental methods, such as 2' hydroxyl acylation analysed by primer extension (SHAPE), make use of the highly reactive 2' hydroxyl group to probe and predict RNA secondary structure experimentally. The nitrogenous bases chemically bind to the hydroxyl group at 1' position of the ribose to form a *nucleoside*. Nucleosides are typically classified into two groups, purines and pyrimidines.

To form a nucleotide, the 5' carbon of the ribose needs to covalently bind with a phosphate group. In their simplest form nucleic acids are polymers of nucleotides chemically bound by a phosphodiester bond, as illustrated in Figure 2.2.

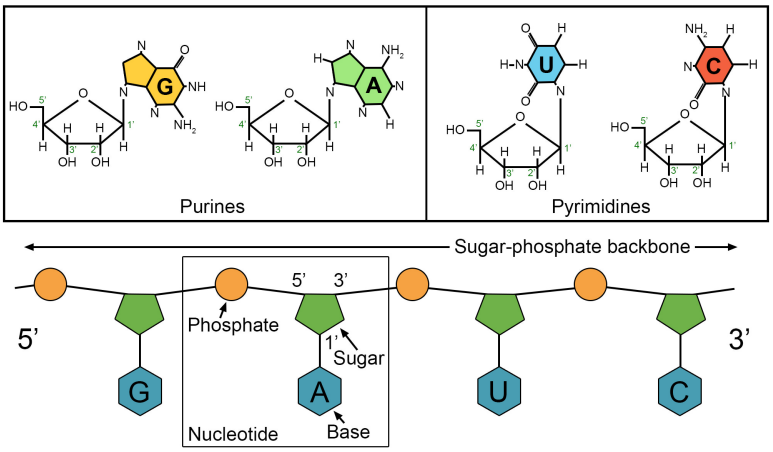

Fig. 2.2 *Nucleotides* - (Top) Purines and pyrimidines, a ribose sugar chemically linked to a nitrogenous base forming a nucleoside. (Bottom) RNA nucleotide chain.

Nucleotides are fundamental building blocks of genetic information. They are also fundamental to the thermodynamics of all living organisms because of the vital role they play in metabolism or energy transfer in cellular organisms. Nucleosides can bind to more than one phosphate group. A nucleoside bound to one, two or three phosphate groups is known as a *nucleoside mono-phosphate* (NMP) or nucleotide, *nucleoside di-phosphate* (NDP) and *nucleoside tri-phosphate* (NTP). The strong covalent bonds formed between phosphate groups store energy and when these bonds are destroyed by enzymes the energy is released.

The bases, (A, C, U, G), found along an RNA strand can chemically bind together, primarily by hydrogen bonds, to form a base pairing. That is, the strand folds in on itself allowing two nitrogenous bases to chemically bind

together. The most common pairings are Watson-Crick complementary, (A-U) and (G-C) pairs. In addition to the Watson-Crick pairs, highly conserved, (G-U) or wobble pairs frequently occur in RNA.

Definition 2.1. *A base pair* (i, j) *is said to be canonical if it is Watson-Crick complementary or a wobble pair, i.e.* $(i, j) \in \mathbb{L}$ *where* $\mathbb{L} = \{(G,C),(C,G),(A,U),(U,A),(G,U),(U,G)\}$.

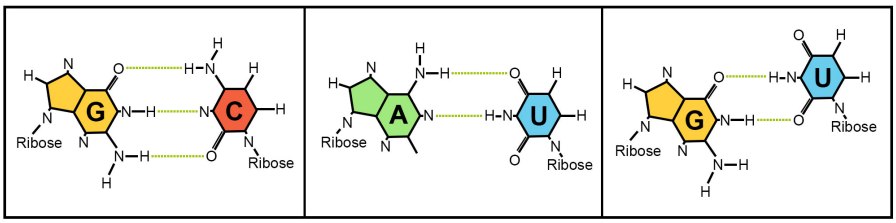

Fig. 2.3 *Canonical base pairs* - Hydrogen bonds formed by the three types of RNA base pairing. From left to right: Guanine (G) - Cytosine (C), Adenine (A) - Uracil (U) and Guanine (G) - Uracil (U).

Each of the three possible base pairings found in RNA have different binding strength based the number of hydrogen bonds formed. Figure 2.3, shows the chemical geometry of the three pairings, collectively referred to as canonical pairs. Watson-Crick (G-C, A-U) pairings have almost identical geometry. The only difference is the (G-C) pairing has three hydrogen bonds and the (A-U) pairing forms two hydrogen bonds. (G-U) pairings are commonly referred to as “weak pairs” or “wobble pairs”. (G-U) pairs like (A-U) pairs are binded by two hydrogen atoms. However, the geometry of the (G-U) pair is very different to that of the Watson-Crick pairs. The skewed geometry of (G-U) pairs make it a weaker pairing in comparison to the Watson-Crick pairs. However, (G-U) pairs are important in RNA structure. (G-U) pairs can stabilise RNA structure [13] and their skewed geometry is believed to play an important role in facilitating the interaction of other molecules such as proteins and other RNAs [14, 15]. Therefore, (G-C) pairings are the strongest possible pairs, followed by (A-U) pairs and then (G-U) pairs.

2.2 Messenger RNA

Historically, RNAs has been viewed as simple and relatively uninteresting ‘messenger carrying’ molecules. Copying information from a gene encoded by

DNA and then being translated into a protein by *ribosomes*. Often depicted by a model first proposed by Francis Crick in 1958 known as the *central dogma of molecular biology*, see Figure 2.4 [16], and restated in 1970 [17]. The dogma states that DNA is *transcribed* into a RNA molecule which then gets *translated* into a protein molecule. Meaning, a gene is copied or transcribed into RNA via a process known as transcription, and once the RNA has been transcribed it is translated into a protein. In this flow of information process, more commonly referred to as gene expression, RNA is simply a messenger carrying molecule. It 'carries' the instruction or message needed to make a protein and is therefore known as a *messenger RNA* (mRNA) or protein-coding RNA.

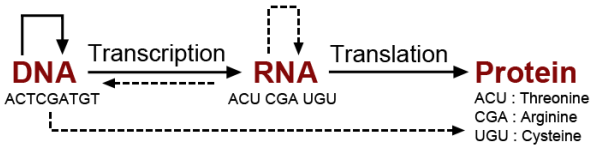

Fig. 2.4 *The Central Dogma of Molecular Biology.* Dashed arrows show transfer of information discovered after the model was restated in 1970.

A mature mRNA transcript consists of two different regions, namely a protein-coding region and a 3' and 5' *untranslated region* (UTR). The protein-coding region contains the sequence needed to make the actual protein molecule and the 3' and 5' UTRs are regulatory regions involved in for example translation and gene expression regulation. mRNA transcription can generally be described by a three step process (see Figure 2.5):

1. **Pre-mRNA and 5' capping**
   Transcription begins by binding of the enzyme *RNA polymerase* (RNAPs) to the regulatory 5' promoter region of a gene. The function of RNA polymerase is to temporarily separate the two strands of a local region of DNA. And, thereby read a sequence of nucleotides. A base which is complementary to a base read by RNA polymerase is then added to the newly formed RNA strand at the 5' end. The addition of bases at the 5' end ensures the RNA sequence is a copy of the sequence on the opposite strand. During transcription a cap is added to the transcribed 5' end of the precursor mRNA. The 5' cap protects the mRNA from cleavage by enzymes; allows for the mRNA to be exported to the cytoplasm and helps initiate protein synthesis.

2. **Polyadenylation**
   Once the gene has been transcribed approximately 100 - 250 adenine (A) bases are added to the 3' end of the mRNA, known as a poly-A tail. A pre-mRNA is composed of coding (introns) and non-coding (exons) regions. The intron regions contain the information required to make a protein.

3. **Splicing and mature mRNA**
   The pre-mRNA is then spliced by a ribonucleoprotein complex called a spliceosome to produce a mature mRNA. Splicing removes from pre-mRNA non-coding introns and joins together exons. The first and last exons (exons 1 and 4 in figure) have special significance and are known the 3' and 5' untranslated region (3'/5' UTR). The 3' and 5' UTR are regulatory regions allowing for sequence specific interaction with other RNAs and proteins. Joining of different exons allows for different mature mRNA isoforms to be produced from a pre-mRNA.

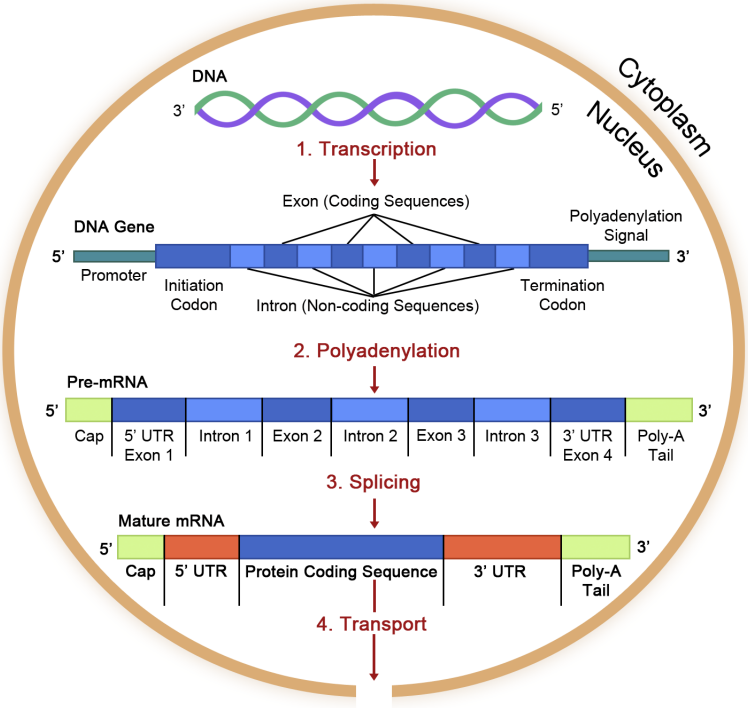

Fig. 2.5 *Messenger RNA Transcription.*

During the 1950s two RNAs were discovered which do not code for a protein, namely *transfer RNA* (tRNA) and *ribosomal RNA* (rRNA), contradicting the simplistic dogma model. It was not until the discovery of catalytic RNAs,

more commonly referred to as ribozymes, in the early 1980s by Thomas Cech's lab who was studying RNA splicing in ciliate protozoan *Tetrahymena Thermophila* [18]. And, Sidney Altman and Norman Pace labs who were jointly studying the *Ribonuclease P* (RNase P) complex [19] that the functional possibilities of RNA were beginning to be realised. Prior to their discovery catalytic activity was believed to be exclusively a function of proteins. Then in 1986 Thomas Cech showed that RNA itself can self-replicate without the need for proteins [20]. The diverse functional capabilities of RNAs as catalytic, self-replicating, information carrying molecules has led to the hypothesis that RNA may be a primordial molecule, known as the *RNA World Hypothesis* [21].

2.3 RNA structure

In cells, DNA most frequently occurs as a double stranded molecule. Where two complementary strands of bases chemically bind together forming a twisted double helix structure. In comparison, RNA most frequently occurs as a single stranded molecule. Allowing RNA to fold into more complex shapes than that of the DNA double helix. Given below are definitions of the three levels of structure used to describe an RNA molecule.

2.3.1 Primary structure

The *primary structure* is the sequence of nucleotides, typically read from the 5' to 3' end.

Definition 2.2. *The primary structure of an RNA molecule is an ordered sequence of characters* $L = (N_1, \ldots, N_n)$ *where* $N \in \{A, C, G, U\}$ *and n is the length of the RNA molecule.*

2.3.2 Secondary structure

The secondary structure of RNA is the folded form of the nucleotide chain, i.e. a 2-dimensional representation after hydrogen bonds are formed between complementary nucleotides. Secondary structure is thus a set of base pairs (i, j) of sequence positions i and j that can form hydrogen bonds. Therefore, in formal terms a secondary structure can be considered as a node-labelled, undirected graph.

Definition 2.3. *An RNA secondary structure is a connected graph $G = (V,E)$, where the set of vertices $V = \{1,...,n\}$ correspond to nucleotides and edges E correspond to the backbone and base pairs formed by hydrogen bonds, $E \subseteq V \times V$ and $L(V) = \{A,C,G,U\}$. The graph G can be represented by an adjacency matrix A with entries $a_{i,j} \in E$ such that [22]*

1. $a_{i,i+1} = 1$ *for* $1 \leq i \leq n-1$
2. *For each i,* $1 \leq i \leq n$*, there is at most one* $a_{i,j} = 1$ *where* $j \neq i \pm 1$ *if* $L(i)$ *and* $L(j)$ *comply with canonical pairs, see 2.1.*
3. *If* $a_{i,j} = a_{k,l} = 1$ *and* $i < k < j$ *then* $i < l < j$.

Conditions 1 - 3 state the following: (1) the phosphate backbone is part of the graph, (2) each nucleotide can only pair with at most one other nucleotide and that pairing must be canonical, i.e. either Watson-Crick or a wobble pair and (3) only nested pairings are allowed, i.e. the graph must be outerplanar without pseudoknots.

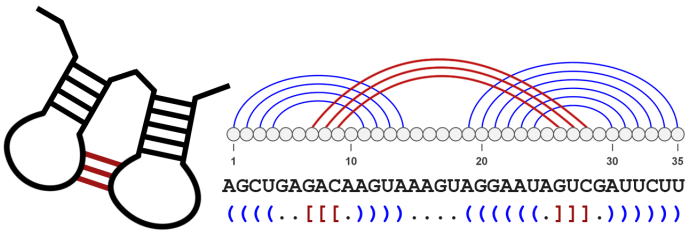

Fig. 2.6 *Kissing hairpin pseudoknot.*

A pseudoknot is a special type of substructure that occurs when a hairpin forms base pairings with another unpaired region creating a knotted structure, Figure 2.6 shows a typical pseudoknot structure known as a kissing hairpin. Most secondary structure prediction tools ignore pseudoknots to reduce complexity of prediction, discussed more in 3.2.5. In addition to the above conditions RNA does not allow for sharp folds, i.e. $\forall(i,j)$ and $i < j \rightarrow i+3 < j$. Therefore, a secondary structure is valid for a given sequence if (1) only binary base pairs are formed, (2) no pseudoknots are present, and (3) each hairpin has at least three unpaired nucleotides. It is important to note that there is a more restrictive definition of secondary structure where in addition to the previous rules no isolated pairings are allowed. These structures are referred to as canonical secondary structures and are assumed to be more stable than a structure consisting of sporadic base pairs.

Definition 2.4. *A secondary structure is canonical if in addition to the conditions defined in 2.3 each pairing has at least one neighbouring base pair, $\forall(i,j)$ there exists a pair (k,l) such that $k = i+1$ and $l = j-1$ or $k = i-1$ and $l = j+1$.*

2.3.3 Tertiary structure

The tertiary structure is the complete three dimensional representation of RNA with many twists and bends; achieved by further hydrogen base pairings and metal ions which help to further stabilise the structure, see Figure 2.7. Ultimately, it is this level of representation that is most desired as it will provide the most insight into function and interaction with other molecules. In comparison to protein structures RNA has a greater degree of freedom and thus predicting the tertiary structure presents many challenges in computational geometry and graph drawing.

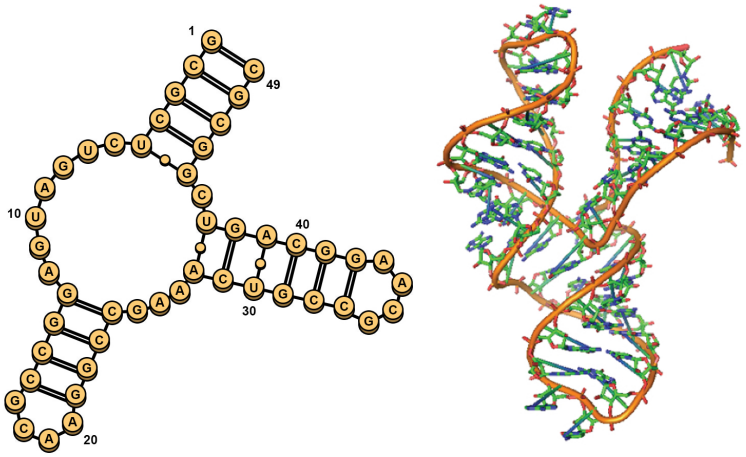

Fig. 2.7 *RNA hammerhead ribozyme* - (Bottom) The primary structure or RNA sequence. (Left) The secondary structure as predicted by, `Vienna RNA package, RNAfold` tool. (Right) the tertiary structure or three dimensional structure based on fluorescence measurements [23] and generated using `PyMol` [24]

RNA folding is thus driven by strong intramolecular forces. And, whilst it is the tertiary structure that is most desired there is much to be gained from understanding RNA at the secondary structure level. Generally, RNA folding is hierarchical, i.e. the secondary structure is formed before tertiary interactions [25]. RNA secondary structure is evolutionary conserved and therefore

includes information on function [26]. Furthermore, evidence suggests that it is at the secondary structure level where most free energy comes from with the tertiary structure contributing only minimally to the stability of the molecule. This is in contrast to protein folding where most free energy comes from the tertiary structure [27, 28]. Therefore, it is believed that the tertiary structure has little impact on the overall major structural shape of RNAs.

2.3.4 RNA secondary structure representation

There are several ways to represent the secondary structure of RNAs. The most simple and easily useable representation is the dot and bracket notation where base pairs are represented as a matching pair of parenthesis.

1. If position i is unpaired then $S_i =$ '.'.
2. If p and q form a valid base pairing and $p < q$ then $S_p =$ '(' and $S_q =$ ')'.

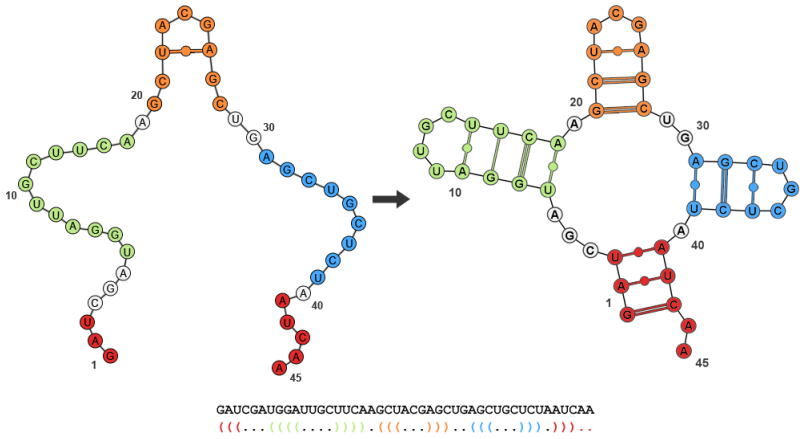

Fig. 2.8 *RNA secondary structure dot-bracket representation.*

2.4 Non-coding RNAs

The latest major RNA discovery, mainly since 2000 onwards, is the existence of a complex network of regulatory non-coding RNAs. These regulatory RNAs are involved in a wide range of celluar processes and shown to play critical roles in many disease causing pathways.

In light of RNAs diverse functional capabilities they are typically categorised into two major groups, protein-coding RNAs known more commonly as messenger RNAs (mRNAs) and non-coding RNAs. The non-coding are further categorised into two groups based on their biological function. Those having a well-defined function are referred to as housekeeping RNAs and those which regulate biological processes are referred to as regulatory RNAs, see Figure 2.9.

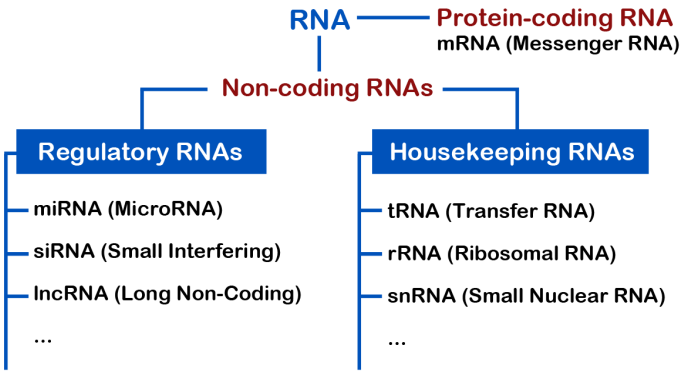

Fig. 2.9 *Non-coding RNAs* - General classes of non-coding RNAs.

Housekeeping RNAs include the most well-known and studied non-coding RNAs such as tRNA and rRNA that are essential for protein synthesis. The regulatory RNAs can further be divided into two subclasses, short non-coding less than 300 nucleotides and long-non-coding RNAs. One class of non-coding RNAs that has received special attention are *microRNAs* (miRNAs) and *small interfering RNAs* (siRNAs).

2.4.1 MicroRNAs and disease

MicroRNAs are a class of short RNAs ($\approx 20-23$ nucleotide) that post-transcriptionally regulate gene expression. MicroRNAs originate from primary transcripts (pri-miRNAs) forming distinctive hairpin structures that are cleaved by the ribonuclease *Drosha*, forming ≈ 60 nucleotide long pre-miRNAs. The pre-miRNAs are further trimmed by the ribonuclease *Dicer* resulting in imperfect duplexes. One strand of the duplex is incorporated into *RNA induced silencing complex* (RISC) to become a functional single-stranded microRNA. microRNAs regulate gene expression in humans by imperfect intermolecular bonding of nucleotides, typically within the 3' UTR, of a target mRNA blocking the ribosome from translating the mRNA into a protein, see Figure 2.10.

See [29] for an in depth discussion on the complexity of microRNA biogenesis. The computational aspects and challenges of predicting miRNA-mRNA interactions are discussed in Chapter 3.

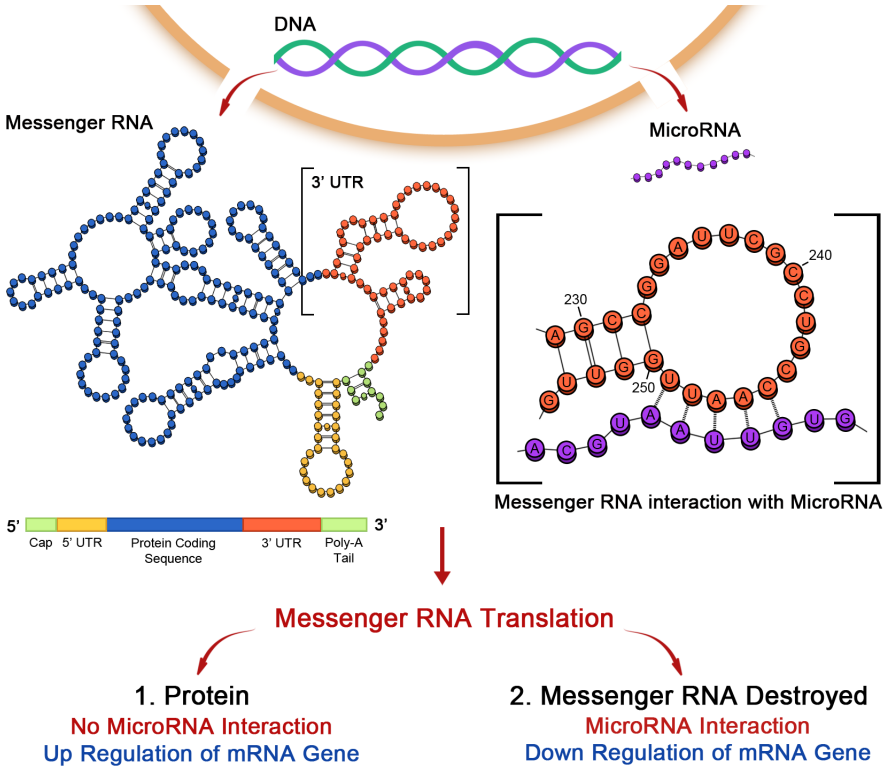

Fig. 2.10 *Messenger RNA Translation* - If a miRNA binds to the 3' UTR of a mRNA then the ribosome becomes blocked from reading all the protein coding sequence. The poly-A tail of the mRNA, acting as a timer, shortens in length over time. Eventually no poly-A tail will exist which signals the release of the ribosome and the mRNA is destroyed.

The first microRNA, *lin-4*, was discovered in 1993 by Ambros *et al.* in *Caenorhabditis elegans* (C. elegans) [30]. However, it was not until 2000 that the second miRNA, *let-7*, was discovered [31] and shortly thereafter a whole class of miRNAs across multiple organisms [32]. As of 2014, 2,588 mature human miRNAs have been characterised (miRBase, release 21) [33] and GENCODE have identified 9,882 small non-coding RNA genes (Ensembl 83, version 24) [4], suggesting the function of many have yet to be discovered. miRNAs are an interesting class of non-coding RNAs because they are involved in many essential biological processes such as gene expression, proliferation, differentiation and apoptosis. And, there is growing evidence they play a critical role in the development and progression of a wide variety of diseases

such as acquired immunodeficiency syndrome, asthma, influenza, hepatitis, measles, Alzheimer's and cancers.

The most well studied microRNA and disease association is cancer. Every day nucleotides of our DNA become damaged and can result in a permanent mutation, e.g. a guanine base being changed to a cytosine. This damage can be the result of both environmental and normal biological processes. For example, damage to the genome of lung cells can occur from tobacco smoke and UV-radiation from sunlight can damage skin cells. It is estimated that the genome of each human cell acquires $\approx$ 20,000 damages every day [34]. Normally, DNA damage is recognised and repaired. However, sometimes the damage is not recognised or the repair procedures make a mistake resulting in a permanent mutation of a nucleotide. It is estimated that on average 1 mutation a day survives in human cells [34]. Over an individual's lifetime these mutations build up and increase their chance to develop cancer. For example, if a mutation occurs within a gene that produces the proteins essential for recognising, controlling or carrying out DNA repair. Our genome contains genes that both promote, called proto-oncogenes such as *Human Epidermal Growth Factor Receptor 2* (HER2), and suppress cell proliferation called tumour suppressors such as *Anaphase Polyposis Coli* (APC) and *Tumor Protein p53* (TP53). Under normal conditions these genes work in harmony balancing the effects of each other. If something disrupts the normal regulation of these genes then proliferation can become deregulated. If a proto-oncogene is too active then it can lead to cancer. Likewise, if a tumour suppressor becomes too inactive then cells will proliferate too rapidly.

The link between microRNA and cancer was first proposed by Calin *et al.* in 2002 when it was discovered that *miR-15* and *miR-16* are down-regulated in chronic lymphocytic leukemia, suggesting microRNAs may act as tumour suppressors [35]. However it took until 2005 to find proof that microRNAs can be both oncogenic, called oncomiRs, and tumor suppressive when Johnson *et al.* found *let-7* targets the 3' UTR of the onocogene *RAS* [36]. Oncomirs negatively regulate tumor suppressor genes resulting in uncontrolled cell proliferation. There are now numerous studies associating microRNAs with cancer, see [37] and [38] for a recent review on the subject.

It has also been shown that microRNA expression profiles classify human cancer types suggesting microRNAs could be used as both a diagnostic and prognostic biomarker [39]. In 2010, it was reported that overexpression of a single microRNA is sufficient to cause cancer [40]. Due to the growing number

of studies a number of databases, such as `HMDD`[41] and `miRGator`[42], have been setup to track experimental microRNA disease associations. For this reason, microRNAs are of special interest in the biomedical field because of their potential use in medicine [43].

Non-coding RNA pharma-therapeutics is a very active area of research with a growing number of clinical trials and increasing number of granted patents [44]. The first experimental RNA candidate drugs to enter clinical trials *Miravirsen* for hepatitis C virus (HCV) is a short antisense RNA drug that inhibits *miR-122* which is hijacked by the virus for replication [45]. This clinical trial demonstrated promising results and was extended into a phase 2 study. However, it was shown that long-term use results in resistance due to mutations in the HCV genome [46]. And in 2013 the first experimental microRNA candidate drug, MRX34, targeting cancer entered into phase 1 of clinical trials [47]. MRX34 mimicks the microRNA tumor suppressor *miR-34* that targets many oncogenes and is found to be down regulated in many cancers. Unfortunately, the clinical trial was halted at phase 1 due to adverse side effects. Furthermore, in 2013 the United States *Food and Drugs Association* (FDA) approved an antisense drug, *mipomersen*, for a genetic condition called *heterozygous hypercholesterolemia* [48]. Mipomersen works by binding to mRNAs that produces apolipoprotein B messenger RNA to inhibit its translation. Whereas traditional drugs typically directly target a problematic protein, RNA therapeutics acts at the stage before proteins; i.e. they aim to silence or inhibit the production of problematic proteins or replace microRNAs that are under-expressed. Clearly, the potential of RNA based therapeutics are far from understood due to many the challenges, such as drug delivery and safety, that must be overcome.

One of the many challenges slowing progress in this field is a lack of knowledge on RNA structure formation and identification of interactions between RNAs and other molecules, such as mRNA-miRNA and RNA-protein interactions. In the microRNA case identifying target mRNAs is challenging because each microRNA typically targets multiple mRNAs. Gene specific identification and verification of microRNA-mRNA interaction can be achieved experimentally. The most popular technique involves cloning the 3' UTR sequence of interest into a firefly luciferase reporter gene assay and measuring fluorescence activity in transfected cells, see [49]. This technique is considered most robust because the experiment can be repeated with a mutated 3' UTR sequence, i.e. at a suspected binding location in order to verify the interac-

tion and identify binding site. Luciferase experiments are typically done in combination with Northern blot analysis or quantitative PCR to test for microRNA and target mRNA co-expression. More recently, several transcriptome wide next-generation sequencing based approaches have been proposed: (1) photoactivatable ribonucleoside–enhanced crosslinking and immunoprecipitation (PAR-CLIP) [50], (2) high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) [51], and (3) crosslinking, ligation, and sequencing of miRNA-RNA hybrids (CLASH) [52]. In general, these high-throughput techniques work by detecting RNA-binding protein sites. More specifically in the case of miRNA-mRNA detection the protein Argonaute which is an essential component of the RNA induced silencing complex. Several verified miRNA-mRNA interaction databases, such as `TarBase 7.0`, `miRTarBase` and `miRecords`, cataloging experimental studies are now available.

Chapter 3

RNA structure prediction, energy landscapes and RNA-RNA interactions

3.1 Introduction

There are two main experimental methods to determine an RNAs secondary structure, enzymatic and chemical probing. Enzymatic probing works by cleaving unpaired regions of structure and chemical probing, such as *2'-hydroxyl acylation analyzed by primer extension* (SHAPE), work by reacting with the backbone of RNA and probing its mobility. Chemical probing techniques are currently the only way to determine RNA structures *in vitro* and *in vivo* at nucleotide level. Experimental techniques commonly used to determine the tertiary structure of proteins, such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Electron Microscopy, are time consuming, technically challenging and expensive processes.

Currently, in comparison to proteins, these experimental methods do not work very well for RNA molecules. This is evident from statistics from the *Research Collaboratory for Structural Bioinformatics* (RSCB) *Protein Databank* (PDB) where, as of December 2016, over 116 thousand protein structures have been submitted and only 1,237 RNA structures, see figure 3.1 for the growth in number of known structures since 1990. Hence the reason why computational RNA structure prediction has been an active area of research since the 1970s. In this chapter we examine the computational techniques to predict RNA secondary structures. The remainder of this chapter is organised

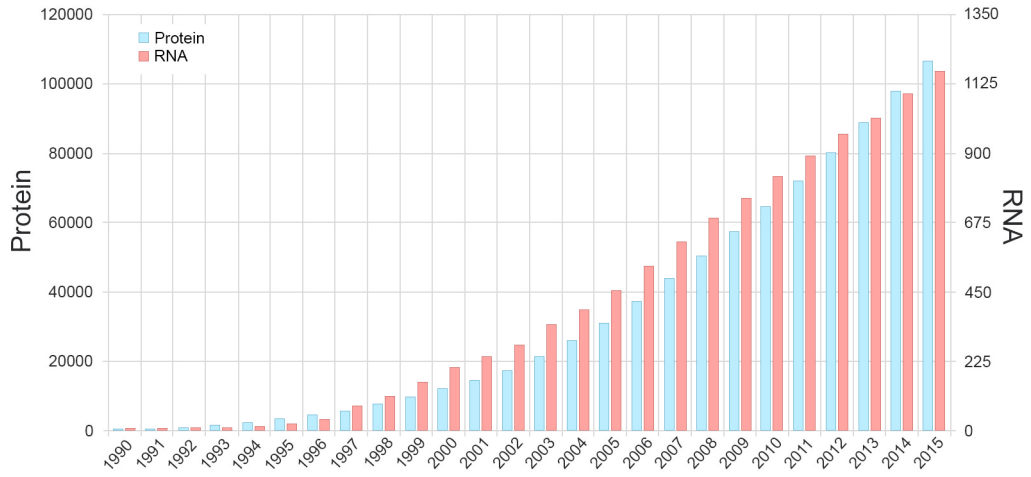

Fig. 3.1 Growth in number of protein and RNA structures deposited to the Protein Databank.

as follows: in Section 3.2 we present an overview of single secondary structure prediction algorithms; in Section 3.3 we explore literature on RNA energy folding landscapes; and in Section 3.4 we discuss literature on the problem of predicting the interaction of two RNA molecules. It is important to note that all the methods discussed in this chapter ignore pseudoknot structures unless explicitly stated otherwise.

3.2 Single secondary structure prediction algorithms

In general, there are three main computational frameworks to predict RNA secondary structure: (1) comparative, (2) probabilistic and (3) energy minimization. Comparative based frameworks use multiple sequence alignment techniques to predict consensus structures, i.e. an optimal structure for a set of sequences. This approach is the gold standard in terms of prediction accuracy and has been successfully applied to ribosomal RNA sequences. This approach is taken for example by the tool `Pfold` which make use of a Stochastic Context Free Grammars (SCFGs) and phylogenetic trees in the prediction [53].

A major limitation of comparative frameworks is they require sets of *homologous sequences*. For many RNAs, especially lncRNAs, it is very difficult to find a set of sequences showing strong conservation. Because of low sequence conservation Sankoff proposed a model that can simultaneously fold and align RNAs sequences [54]. Probabilistic frameworks employ statistical machine

learning techniques to estimate folding parameters from a set of known secondary structures. Several probabilistic frameworks have been proposed of which most apply SCFGs. The most known is `ContraFold` which is based on conditional random fields [55]. The most popular framework is energy minimization using experimentally determined molecular thermodynamic data.

3.2.1 Recursive decomposition of secondary structure

Computational RNA structure prediction dates back to 1978 when Waterman proposed a recursive decomposition scheme to enumerate the number of secondary structures [22]. The underlying idea of this decomposition scheme is as follows: any substructure, on the interval $s[i,\ldots,j]$, can be further divided into smaller substructures by examining if a position i is paired or unpaired.

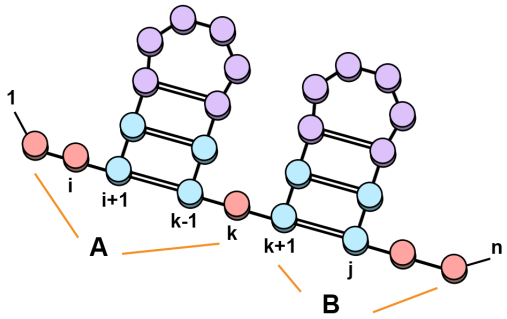

Fig. 3.2 Waterman's structure decomposition.

If i does not pair then the substructure consists of the base pairings in $s[i+1,\ldots,j]$. However, if i is paired with another base k such that $i < k \leq j$ then base pair (i,k) subdivides the structure $s[i,\ldots,j]$ into two smaller parts, namely $s[i+1,k-1]$ with base pair (i,k) closing the substructure and adjacent to it a substructure $s[k+1,j]$. Therefore, if a pairing is formed the structure on the interval $s[i,\ldots,j]$ is equal to substructure $\mathrm{A} = s[i+1,\ldots,k-1]$, the pairing (i,k) and substructure $\mathrm{B} = s[k+1,\ldots,j]$, see figure 3.2.

Secondary structure loop motifs

Using this decomposition scheme it is possible to define secondary structure in terms of different types of substructures or loops [22, 56]:

- **Stacked pair:** consists of two consecutive base pairs, i.e. if i and j are paired and $i+1$ and $j-1$ are paired. Consecutive stacked pairs form helical stem loops, i.e. $(i,j),(i+1,j-1),(i+2,j-2),\ldots$.
- **Hairpin loop:** consists of one closing base pair (i,j) and $m \geq 3$ unpaired bases, i.e. i and j are paired and $j-i>3$, and k are unpaired bases such that $\forall k, i<k<j$.
- **Interior loop:** consists of one closing base pair, a single enclosed base pair and unpaired bases between the base pairs. Formally, for base pairs (i,j) and (k,l) such that $i+1<k<l<j-1$ and unpaired bases m, $\forall m, i<m<k$ and $j<m<l$.
- **Bulge loop:** a special type of internal loop where there are no unpaired bases on one side.
- **Multiloop:** a loop which has one closing and at least two enclosed base pairs. For base pairs $(i_1,j_1,i_2,j_2,\ldots,i_m,j_m)$ where $m \geq 2$ and $i_1<i_2<j_1<\cdots<i_m<j_m<j$ and unpaired bases k, $\forall k, i<k<i_1, j_1<k<i_2,\ldots,j_m<k<j$.
- **Exterior loop:** a special case since it does not have any closing base pair. An exterior loop consist of all paired and unpaired bases that are not part of any other loop, they are unpaired bases between other loops. Any unpaired bases at the 3' and 5' ends are called *dangling ends*.

Figure 3.3 shows the different types of loop structures as defined above.

Number of secondary structures

Using the decomposition scheme Waterman *et al.* derived the following recursion to enumerate the number of secondary structures: Let S_n be the total number of secondary structures and $S_0 = S_1 = S_2 = 1$. Then for $n > 2$, S_n satisfies the following recursive formula [57]:

$$S_{n+1} = S_n + \sum_{k=0}^{n-2} S_k S_{n-k-1} \tag{3.1}$$

If a new base added to S does not pair then the total number of structures is equal to the number for S_n. If the new base pairs with k then the total number of structures is the product of the number of substructures S_k and S_{n-k-1}, e.g.

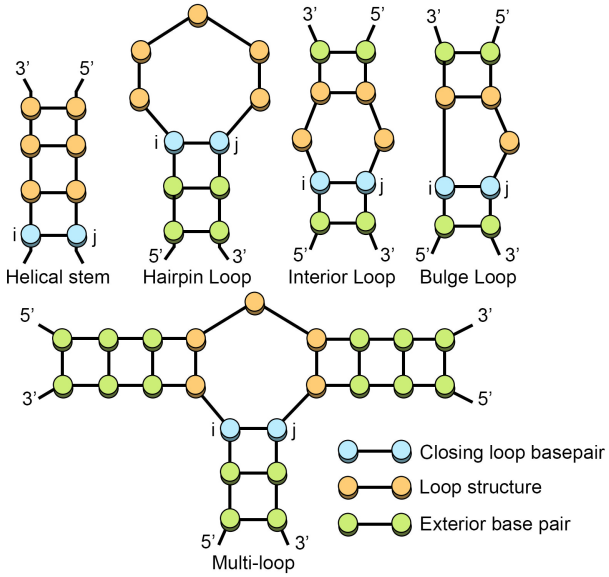

Fig. 3.3 Secondary structure loop decomposition.

all valid substructures of region A by all valid substructures of B in figure 3.2. From the above recursion Zuker derived, using a stochastic approach, an asymptotic value for the recursion $S_n \approx 1.8^n$ [58]. Considering the average length of a human 3' UTR region sequence is several hundred nucleotides exhaustive enumeration is impossible.

3.2.2 Nussinov-Jacobson base pair maximization

Nussinov and Jacobson proposed in 1978 a dynamic programming algorithm to compute the structure with maximum number of base pairs [59]. Nussinov and Jacobson apply the decomposition scheme proposed by Waterman, such that for a substructure $s[i,\ldots,j]$ if i is unpaired then the maximum number of base pairs is the same as substructure $s[i+1,\ldots,j]$. However, if i is paired to base k then the maximum number of base pairs is equal to the sum of substructure $M_{(i+1,k-1)}+1$ for base pair $(i,j)+$ substructure $M_{(k+1,j)}$. In simple terms, the substructure $s[i,j]$ maximizes the number of base pairs if its decomposition leads to the substructures with maximum number of base pairs.

$$M_{(i,j)} = \max \left\{ \begin{array}{l} M_{(i+1,j)} \\ \max\limits_{\substack{i<k\leq j \\ (i,k)\in \mathbb{L}}} M_{(i+1,k-1)} + M_{(k+1,j)} + 1 \end{array} \right\} \tag{3.2}$$

The first line in the recursion 3.2 is if i does not pair with another base and the second line if (i,k) forms a base pairing splitting the structure into two substructures. At the end of the calculation the maximum number of base pairs can be found at $M_{(1,n)}$. To determine the actual structure with maximum base pairs, $M_{\max}$, a backtracking procedure is required to combine all subsolutions. The structure decomposition scheme proposed by Waterman *et al.* and the dynamic programming approach to maximize base pairs set the foundations for RNA secondary structure prediction. Nussinov's algorithm has a time complexity of $O(n^3)$ because it iterates over the variables i, j and k and a space complexity of $O(n^2)$ for a sequence of length n. The problem with this approach is it assumes each base pair is equally likely to occur. However, a GC pairing is more favourable than a GU pairing because of the number of hydrogen bonds formed.

3.2.3 Thermodynamics of RNA secondary structures

In 1980, Nussinov and Jacobson modified their algorithm to incorporate a basic scoring function [60]. The recursion for this algorithm is identical to their maximum matching version except now the problem is a minimization problem. And, each type of base pairing is scored depending on the number of hydrogen bonds formed as shown in 3.4.

$$M_{(i,j)} = \min \left\{ \begin{array}{l} M_{(i,j-1)} \\ \min_{\substack{i \le k \le j-1 \\ (k,j) \in \mathbb{L}}} M_{(i,k-1)} + M_{(k+1,j-1)} + E(k,j) \end{array} \right\} \tag{3.3}$$

where

$$E(k,j) = \begin{cases} -1 \text{ if } \{(G,U),(U,G)\}, \\ -2 \text{ if } \{(A,U),(U,A)\}, \\ -3 \text{ if } \{(G,C),(C,G)\} \end{cases} \tag{3.4}$$

However during the 1970s, biochemical analyses on the stability of RNA binding lead to a hypothesis that the stability of a RNA structure is derived from both base pair types and base pair stacking interactions [61, 56]. In other terms, the hydrogen bonds formed by a base pairing and from interaction between adjacent base pairs. Therefore, the simple scoring scheme used by Nussinov *et al.* is too simplistic to accurately predict secondary structures.

Gibbs free energy and the nearest neighbour energy model

Thermodynamic prediction assumes that at equilibrium RNAs fold to a unique functional structure of lowest free energy change known as the Minimum Free Energy (MFE) structure. The free energy change, ΔG measured in kilocalories per mol (kcal/mol) quantifies the difference in free energy between a unfolded RNA chain and a folded state. Gibbs free energy, G, is the total amount of energy available to do work within a defined system, e.g. a cell. It defines the direction of a spontaneous change of a chemical process from a non-equilibrium state to equilibrium at constant pressure and temperature. Gibbs free energy is defined by the following function:

$$\Delta G = \Delta H - T\Delta S,$$

- ΔH **- Enthalpy Change**
  Enthalpy is the total energy or heat exchange between the system and its surrounding environment. The energy absorbed or released from hydrogen bonds.

- ΔS **- Entropy Change**
  Entropy is a thermodynamic function which measures the disorder of a system. For example, consider the two states of water ice and steam. In the ice state the molecules form a well organised lattice arrangement. Whereas in the steam state the molecules are much more freedom and are therefore much more unpredictable. Steam has higher entropy and is in a more disordered state than ice. In RNA folding entropy can be viewed as the energy required for the RNA to organise itself into a folded stable state. A positive ΔS value indicates an increase in disorder and a negative value a decrease in disorder.

- T **- Temperature in Kelvin.** Typically RNA structure prediction is modelled at 37 degrees Celsius.

In general, negative free energy states are favourable and positive free energy states unfavourable. By the late 1980s, extensive work had been completed on tabulating free energy values from melting point experiments on RNA substructure loops [62, 63]. This work resulted in a more biologically realistic thermodynamic model known as the *Nearest Neighbour Model*. There are three general points to be aware of with this energy model:

1. It is assumed that free energy of a structure is equal to the sum of its loops, i.e. the free energy contribution of a loop structure is independent from all other loops.

2. All substructure loops except for helical stem loops are unpaired and are in free energy terms typically destabilising. Of course any helix must have at least one unpaired loop that can contribute a positive free energy value.

3. Not all substructure loop and sequence combinations have been experimentally measured. There are simply to many possible combinations, an unpaired loop of size n has 4^n possibilities. However, it is possible to derive an estimate value for some parameters mathematically [64].

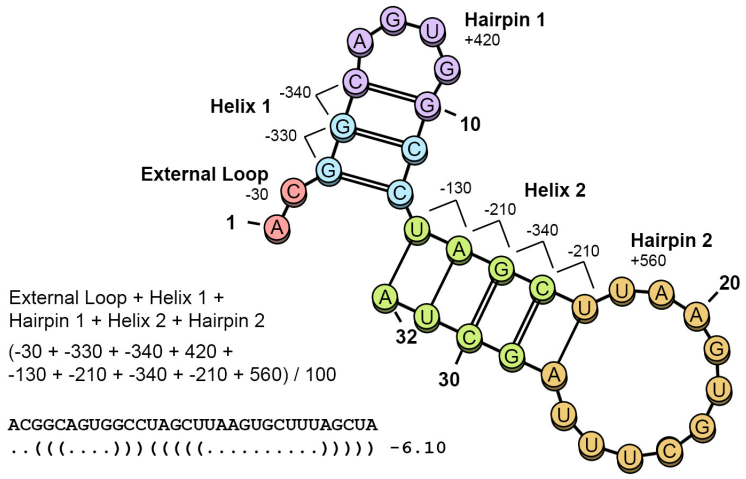

Fig. 3.4 *Nearest neighbour energy calculation:* energy values calculated using `RNAeval` [65].

Figure 3.4 shows an example nearest neighbour free energy calculation. In this example, the external loop (`AC`) at position 1 contributes negative free energy change with value -0.3. Helix 1 consists of two (`GC`) pairs and one (`CG`) pair, where a (`GC`) pair next to another (`GC`) has value -3.3. And, a (`GC`) next to a (`CG`) pair has value -3.4, resulting in helix 1 having a total value of -6.7. Hairpin 1 consists of four nucleotides and contributes a positive value of 4.2 to the overall energy calculation. The energy of the two other loops, helix and hairpin 2, are calculated in a similar way resulting in the structure having a free energy change score of -6.10.

The *Nearest Neighbour Database* (`NNDB`) maintained by Turner and Mathews make publicly available a set of free energy parameters for RNA loop

sequence pairs [66]. This database compiled from published experimental studies dating back to the 1970s consists of hundereds of free energy values, of which were last updated in 2004. And, in 2006 enthalpy values were added for the 2004 free energy values allowing for prediction of structures at temperatures other than the standard 37 degrees Celsius [67]. The parameter set provided by `NNDB` is the most commonly used model incorporated into RNA secondary structure prediction tools. A variant of the Turner *et al.* parameter set is the *Andronescu model* which uses a Constraint Generation method to estimate additional parameter values [64].

3.2.4 Zuker's minimum free energy prediction

Based on the loop decomposition scheme of Waterman and dynamic programming recursion of Nussinov *et al.*, in 1981 Zuker *et al.* proposed a more sophisticated algorithm to predict the structure with minimum free energy change [68]:

$$\mathrm{MFE}(R) = \sum_{S \in \mathsf{C}} E(S) \tag{3.5}$$

where C is the conformation space of sequence R. A key difference of Zuker's algorithm to the Nussinov *et al.* algorithm is it incorporates experimentally determined energy values and distinguishes between helices, hairpins and interior loops. And in 1984 Zuker and Sankoff included multiloop recursion to their algorithm [58]. The extra information computed by Zuker's algorithm requires the computation of four dynamic programming matrices to take into consideration energy contribution of hairpin, multi and interior loops:

1. $F_{i,j}$ to compute and keep track of the minimum free energy between i and j. Equivalent to Nussinov's recursion scheme 3.2.

$$F_{i,j} = \min \begin{cases} F_{i,j-1} \\ \min_{i<k<j} F_{i,k-1} + C_{k,j} \end{cases} \tag{3.6}$$

2. $C_{i,j}$ to compute the minimum free energy when i and j form a base pair. Where H is the energy of a hairpin, J the energy of interior loop and a and b are penalty parameters.

$$C_{i,j} = \min \begin{cases} H(i,j), \\ \min_{i<k<l<j} J_{i,j,k,l} + C_{k,j}, \\ a + b + \min_{i<k<j} M_{i+1,k-1} + M'_{k,j-1} \end{cases} \tag{3.7}$$

3. $M_{i,j}$ to compute the minimum free energy of a multi-loop containing at least one helix, with c being a penalty parameter.

$$M_{i,j} = \min \begin{cases} c + M_{i,j-1}, \\ \min_{i<k<j}(k-i)c + b + C_{k,j} \\ \min_{i<k<j} b + M_{i,k-1} + C_{k,j} \end{cases} \tag{3.8}$$

4. $M'_{i,j}$ to compute the minimum free energy of a multi-loop containing exactly one helix and base i is paired.

$$M'_{i,j} = \min \begin{cases} c + M'_{i,j-1}, \\ b + C_{i,j} \end{cases} \tag{3.9}$$

The asymptotic time complexity of Zuker's algorithm is $O(n^4)$ with a space complexity of $O(n^2)$ for a sequence of length n. The dominating factor in the computation comes from calculation of interior loops. Hofacker *et al.* bound the number of free bases of internal loops to a maximum of 30 nucleotides reducing the complexity to $O(n^3)$ time [69]. Without limiting internal loop sizes, Lyngso *et al.* reduced the runtime complexity to $O(n^3)$ by using additional space $O(n^3)$ [70]. Zuker's MFE based prediction is the most commonly used approach to analyse an RNAs structure. Implementations of this algorithm can be found in for example the tool `UNAFold` [71], *Vienna RNA package* `RNAfold` [65], and `RNAstructure` tool [72]. Eddy reported in 2004 that energy-based secondary structure prediction algorithms correctly predict on average only 50% - 70% of base pairs correctly [73].

3.2.5 Pseudoknot structure prediction

Prediction of MFE pseudoknotted structures is a challenging task, proven to be NP-hard [74]. However, there are many classes of pseudoknot structures of varying complexity. Over the past decades several advances have been made in the development of polynomial time pseudoknot prediction tools. Generally, two approaches have been taken: (1) exact approaches using dynamic programming, and (2) heuristic based approaches. The exact approaches

reduce complexity by restricting the types of pseudoknots included into the algorithm. For example, Rivas and Eddy extended the Zuker algorithm allowing for prediction of a specific type of pseudoknotted structure. With complexity of $O(n^6)$ time and $O(n^4)$ space for a sequence of length n, it is only applicable to sequences of length < 100 nucleotides [75]. Tools which take this approach include `NUPACK` [76], `PKnots` [75] and others. Others propose heuristic approaches such as `HotKnots`, `HFold`, and `ProbKnot` to reduce time complexity.

3.3 RNA energy landscapes

In the previous section, we considered single secondary structure prediction algorithms. More specifically prediction of a structure in thermodynamic equilibrium, a problem with this approach is provides no indication of how likely, for a given sequence, the MFE structure will be folded to. Furthermore, the MFE structure is not unique, i.e. there can be more than one structure having the same free energy value. There are several reasons why considering suboptimal structures close the MFE structure is a good idea:

- **Co-transcriptional folding**
  RNAs start folding as they are being transcribed from DNA, known as *co-transcriptional folding*. Allowing for secondary structure to form at the 5' end of RNA before the 3' end has been fully synthesized. Thus, the RNA must refold for any long range interactions between 3' and 5' end to occur. Transcription rates vary by organism, ranging from ≈ 200 nucleotides per second in phages [77] to $\approx 10-20$ nucleotides per second for human polymerase II, typically with many pauses [78]. It has been reported from experimental studies that the rate of transcription has an impact on RNAs structure and as a result its function [78, 77, 79, 80].

- **Multi-functional structures**
  The MFE structure may not be the functional structure. Evidence suggests RNAs does not always fold into a single, static, structure instead they can fluctuate between different low energy conformations. As stated by Levinthal, in the context of protein folding, a protein cannot randomly search the conformation space for the functional structure on a biological timescale, inferring there must be a preferred pathway to the

functional structure [81]. A well studied example of multiple structures are *riboswitches* which can fold into two distinct structures representing their on and off states [82]. A riboswitch is a specific part of mRNA, typically found within the 5' UTR, that binds a metabolite as a ligand changing structure.

Riboswitches play a crucial roles in regulating gene expression in plants and bacteria such as the human pathogen *Vibrio Vulnificus* which adapts three distinct structures [83]. Another well documented example of multiple structures are *viroids*. Viroids are single-stranded circular RNAs that can cause plant diseases such as *potato spindle tuber* [84]. Solomatin *et al.* report, using single molecular experiments on Tetrahymena group I ribozyme, RNAs to have complex topology of folding landscapes. Leading them to suggest that instead of a single functional state, i.e. the global minimum, multiple pathways leading to multiple functional states may exist [85].

- **Cell environment**
  Changes in cell environment such as temperature change can cause perturbations of RNA secondary structures [83, 86, 87]. RNA thermometers change their structure during heat or cold shock. In other words, it is highly unlikely that, at least for long RNA transcripts, RNA secondary structures are fixated for their lifetime in one particular state.

- **Interaction with other molecules**
  RNAs typically do not fold in isolation. Interaction with other molecules such as metal ions, helicases, chaperones and other RNAs can change structure [88].

- **Free energy parameters**
  Not all free energy values have been experimentally determined and it is likely there are some inaccuracies. Pure thermodynamic prediction only measures stability of the unfolded open chain to an equilibrium state, i.e. it ignores how the RNA folds and any events after a stable state is reached. Additionally, no pseudoknot free energy values are included in the standard Turner nearest neighbour model.

In order to get better insight into RNA structure to improve predictions, by incorporating the complexities described above, algorithms that consider

suboptimal structures need to be considered. And, suboptimal RNA structures are best understood by means of energy landscapes.

Definition 3.1. *The energy landscape,* $\mathsf{L}(R) = [\mathsf{C}, \mathsf{N}, E]$*, of an RNA sequence* R *can be described by three components: a conformation space* C *of secondary structures, a neighbourhood function* N *and a free energy evaluation function* E.

The conformation space C consists of all valid RNA secondary structures for sequence R. One of the first attempts to generate a set of suboptimal structures within a energy increment of the MFE can be traced back to 1989 when Michael Zuker modified his dynamic programming [89]. The modified algorithm calculates for each valid base pair of a given sequence the lowest energy structure containing that base pair, thus outputting at most $n(n-1)/2$ structures. However this approach does not calculate all suboptimal structures as by definition the MFE structure contains the lowest energy base pairs. In 1999 Wutchy *et al.* [90] developed, based on ideas proposed by Waterman and Byers to calculate shortest paths in networks [91] and Zucker's MFE approach, an algorithm to calculate all secondary structures within a user defined energy offset of the MFE conformation. An implementation of Wuchty's algorithm can be found in the *Vienna RNA package*, `RNAsubopt` tool [65]. As the number of structure scales exponentially with increasing sequence length and increasing energy offset this tool can only be applied to relatively short sequences or small offsets. Recently, Stone *et al.* developed a modified version of the Wuchty algorithm for parallelization with additional filters such as the length of helices [92].

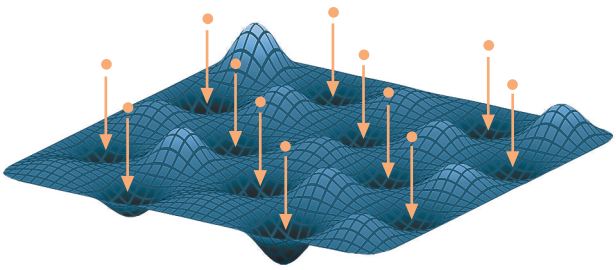

Fig. 3.5 *RNA energy landscape with multiple metastable states.*

The neighbourhood function $\mathsf{N}(\mathsf{S})$ is the set of structures reachable from S by the application of transition operations described by a move set. A move set is a set of operations used to move within in the energy landscape. A move

set must satisfy the following properties: (1) applying a operation results in a structure present in the conformation space, (2) each operation must have a reverse operation, and (3) it must be ergodic, i.e. possible to transition from any structure in C to any other structure in C [93]. The most basic of move sets is the insertion and deletion of single base pairs. An alternative move set for canonical structures (see 2.4) is the following:

1. **Single base pair insertion**
   A single base pair (i, j) may be added if it extends an existing helix and it does not violate the minimum hairpin size.

2. **Double base pair insertion**
   Two consecutive base pairs (i, j) and $(i+1, j-1)$ or $(i-1, j+1)$ may be added if (i, j) does not extend a helix, i.e. positions $i-1$ and $i+1$ are unpaired or $j-1$ and $j+1$ is unpaired. And, the minimum hairpin loop size is not violated.

3. **Single base pair deletion**
   A single base pair may be deleted if it does not result in any isolated base pairing.

4. **Double base pair deletion**
   Two base pairs may be deleted if their deletion does not result in any isolated base pairs.

Flamm *et al.* extend the basic move set to include a shift operation, a base pair deletion followed by a base pair insertion, to model helix diffusion [93]. Using the definition of energy landscapes it is now possible to describe RNA folding as a series of transitions or more formally as a walk on the energy landscape.

3.3.1 RNA energy landscape definitions

Given below are definitions used to describe RNA energy landscapes:

Definition 3.2. *A walk or folding path between conformations* x_1 *and* x_k *is a sequence of conformations* $x = x_1, x_{\ldots,x_k}$ *where* $1 \leq i \leq k : x_i \in \mathsf{C}$ *and* $(1 \leq i \leq k) : (x_i, x_{i+1}) \in \mathsf{N}$. *A walk is said to be a descent walk if* $(1 \leq i \leq k) : E(x_{i+1}) < E(x_i)$ *and a steepest descent walk if* $x_{i+1} = \arg_{x \in \mathsf{N}(x_i)} \min E(x)$.

The terminating conformation of a descent walk is a *local minimum* or *metastable* structure.

Definition 3.3. *A structure s is said to be a local minimum or metastable if* $\forall x \in \mathsf{N}(\mathsf{s}) : E(s) \leq E(x)$.

A structure is therefore a local minimum if **all** its neighbouring structures have higher energy. Similarly, the terminating conformation of a hill climb where all neighbouring structures have lower energy is known as a *local maximum*. Associated with each local minimum is a basin of attraction:

Definition 3.4. *Every conformation belonging to* C *is mapped to a local minimum conformation that can be reached by a descent walk. Every local minimum lm has associated with it a basin of attraction* $B(lm)$ *which is a set of structures that map to lm by application of a descent walk, i.e.* $B(lm)$ *denotes the set of suboptimal structures which by an energy decreasing walk reach local minimum lm.*

RNA energy landscapes are typically rugged, high dimensional surfaces with many mountains and valleys [94, 85]. Folding between any two secondary structures, x_1 and x_k, from a energy landscape can be described as a *direct* walk or folding path. A direct path is the base pair distance between two structures, i.e. a shortest path between the two structures. More formally:

Definition 3.5. *Let* P_{tot} *denote the total number of base pairs for a given structure. A minimum direct path* P_{x_i,x_j} *between two local minima* x_1 *and* x_k *consists of* $P_{tot} = P_S + P_D$ *moves, where* P_S *is the number of shared base pairs between* x_1 *and* x_k *and* P_D *is the number of distinct base pairs. A base pair is distinct if it is present in only one structure. On a shortest path only addition and removal of distinct pairs are allowed, i.e. assuming* x_1 *is the start structure then there will be two sets of moves. A deletion set consisting of base pairs in* x_1 *but not in* x_k *and and insertion set consisting of base pairs in* x_k *but not in* x_1. *[95]*

Taking into consideration the energy of the intermediate secondary structures on a direct path a special case structure called a saddle point can be encountered:

Definition 3.6. *A saddle point s is the structure having maximum free energy on the folding path P. Meaning s has at least two gradient walks to distinct local minima, i.e. a saddle point separates two basins of attraction.*

Knowledge of local minima and saddle point energies allows for a *energy barrier* to be associated with each local minimum:

Definition 3.7. *The energy barrier of local minimum x_1 is the difference in free energy between the lowest saddle point among all possible direct paths and the local minimum, i.e.* $\max E(P_{x_i,x_j}) - E(x_1)$.

Energy barriers are vital information for kinetic or dynamic based folding. Computation of a optimal free energy folding path between two structures is NP-complete [96]. For this reason a number of heuristics have been proposed. One of the first to have considered this problem was Morgan and Higgs who proposed in 1998 a greedy heuristic to determine optimal folding path or energy barrier using the basic Nussinov energy model 3.4 [95]. Energy barriers are important in kinetic folding because they give indication of the energy required for an RNA molecule to refold into a different basin of attraction, i.e. how unstable the structure needs to become by disruption of base pairings before it can fold into a different stable state. Flamm *et al.* proposed in 2001 the tool `findpath`, a breadth-first search with bounded look-ahead heuristic based on the Morgan and Higgs procedure [97]. In 2002 Flamm *et al.* developed the exact algorithm `Barriers` to coarse-grain the energy landscapes into basins of attraction [98].

The algorithm of `Barriers` requires as input a energy sorted list of all sub-optimal structures within some energy interval, e.g. as output by `RNAsubopt`, and outputs the connectedness of the landscape which can be depicted visually by means of a *barrier tree*. Using this tool one can determine the number of local minima, their energy barriers and saddle points. The general idea of the algorithm, often described as *flooding the landscape*, is considering the energy sorted list in increasing order, generate the neighbourhood of a structure if one of its neighbours has been seen before then it is not a local minimum. If none of its neighbours have been seen before then it is a local minimum. By storing all suboptimal structures belonging to each basin separately it is possible to determine saddle points, i.e. structures with a neighbour in more than one basin. Figure 3.6 shows the barrier tree for a sequence consisting of 44 nucleotides: `AGCUAGUGAGGAAAUGUUUUAGACGAUAAAUUUGAGAAAACGCC`.

In 2010 Dotu *et al.* proposed the meta-heuristic approach `RNAtabupath`, where a semi-greedy heuristic and a tabu list of moves is maintained to prevent local search from becoming trapped in local optima [99].

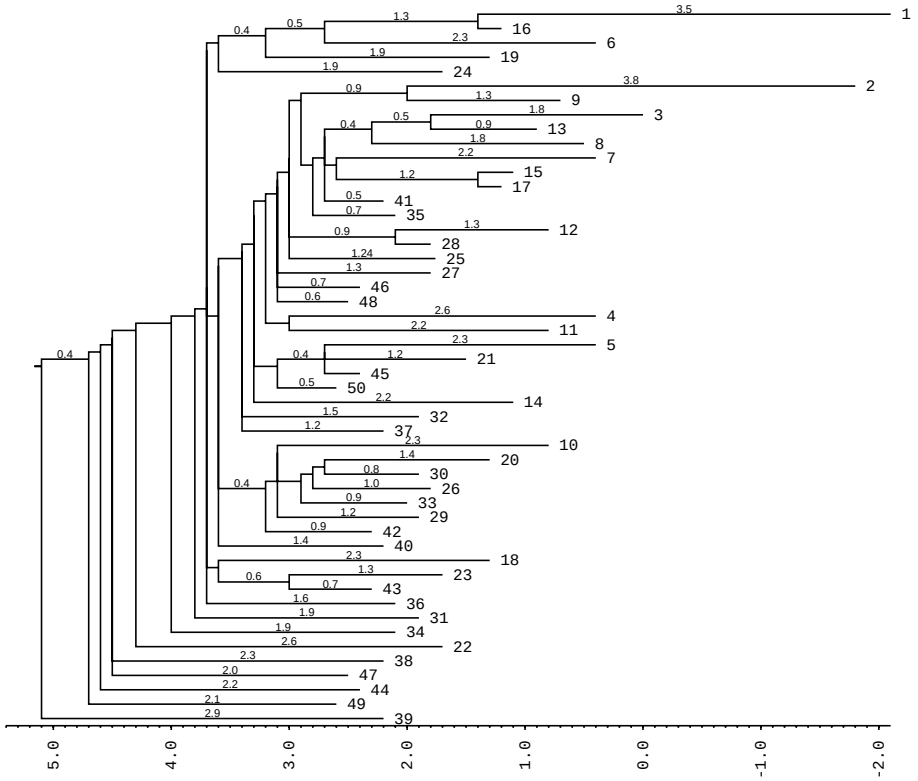

Fig. 3.6 *Barrier tree:* connectivity of the 50 lowest local minima within energy offset 40 kcal/mol. of the MFE. `RNAsubopt` predicts 69,162 structures and `Barriers` 244 local minima with energy barrier > 0.1 kcal/mol.

3.3.2 McCaskill's partition function

In 1990, McCaskill proposed, based on ideas from statistical mechanics, a modified version of Zuker's algorithm to compute the equilibrium partition function of an RNA [100]. Allowing for the computation of equilibrium properties such as base pair probabilities. McCaskill's algorithm requires calculation of the the partition function:

$$Z = \sum_{S \in \mathsf{C}} e^{-(S)/RT} \tag{3.10}$$

where the summation is the free energy of all secondary structures, $-(S)$ is the free energy of structure S, R is the gas constant and T temperature. In simple terms, the modifications made to Zuker's MFE algorithm is the substi-

tution of each minimum to a summation and each addition to a multiplication. The frequency of a particular structure, from an ensemble in equilibrium, is Boltzmann distributed allowing for the probability of a given structure to be computed [100]:

$$P(S) = 1/Ze^{-(S)/RT} \tag{3.11}$$

And, the probability of a specific base pairing [100]:

$$p_{(i,j)} = \sum_{S \in \mathsf{C}} P(S)\delta_{(i,j)}(S) \tag{3.12}$$

where $\delta_{(i,j)} = 1$ if the base pair is in S and 0 otherwise. As this algorithm uses the same recursion scheme proposed by Zuker it maintains the same, $O(n^3)$, time complexity to compute the MFE. Figure 3.7 shows the probability of base pairings, frequency and diversity values for the MFE structure of sequence:

`AGCUAGUGAGGAAAUGUUUUAGACGAUAAAUUUGAGAAAACGCC`

Implementations of McCaskill's algorithm can be found in the tools *Vienna RNA package* `RNAfold`, `Sfold`, `UNAfold`, and `NUPACK` which includes a restricted class of pseudoknots.

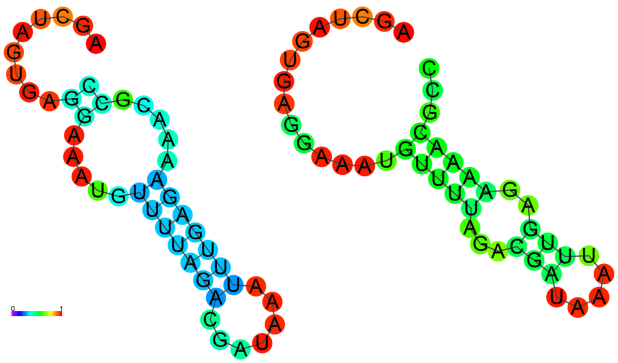

Fig. 3.7 *Base pairing probabilities:* (Left) MFE structure (-1.60 kcal/mol), as computed by the *Vienna RNA package* `RNAfold tool` [65], coloured by base pairing probabilities and for unpaired bases probability of being unpaired. (Right) The centroid structure with base pairing probabilities. Free energy of ensemble = −2.60 kcal/mol. Frequency of the MFE structure in the ensemble = 19.84%, ensemble diversity = 9.69.

In 1996, Cupal *et al.* presented a modification of the McCaskill's algorithm to compute the density of states with a $O(n^5)$ time complexity, i.e. the number

of structures having a user defined energy value [101]. Ding *et al.* proposed a modification of McCaskill's algorithm allowing for statistical representative sampling from a Boltzmann-weighted ensemble [102]. And, in 2005 *et al.* proposed a method to calculate a Boltzmann-weighted representative structure, referred to as a *ensemble centroid structure*. A centroid structure is defined by Ding *et al.* as one having minimum total base pair distance to all structures in the ensemble, also shown in figure 3.7. The authors statistically sample 1,000 structures from their Boltzmann-weighted ensembles and cluster the structures into free energy landscapes based on base-pair distance. For each cluster a centroid structure is calculated and the cluster of which the MFE structure belongs is determined. The authors then compare the cluster centroid structures to the MFE predicted structure. The authors assume the correct structure of their sequences to be the one determined by comparative sequence analysis. They report improved prediction accuracy in comparison to the MFE structure, where the MFE structure falls, in over half of the 81 sequences analysed, outside of the cluster containing the centroid structure that most closely matches the 'known' structure.

3.4 Prediction of RNA interactions

To understand the biological function of a molecule it is vital to understand how it interacts with other molecules. RNA interacts with many other molecules, including small molecules such as amino acids, nucleic acids and proteins. Many biological processes are governed by RNA-RNA interactions. In this section we review computational approaches to predict the site of RNA-RNA interactions, focusing on microRNA target site prediction. Computational RNA interaction prediction typically takes into consideration three main components:

1. Sequence complementarity.
2. Energetics of the formation of intermolecular base pairs.
3. Intramolecular base pairings formed by the secondary structures of interacting molecules.

The simplest approach to predict RNA-RNA interactions is to ignore intramolecular structure and consider only potential intermolecular base pair

formations, i.e. sequence complementarity. This approach is taken for example by the microRNA target site prediction tools `RNAhybrid` [103], `RNAduplex` [65], `RNAplex` [104] and `miRanda` [105]. In general, these tools use a modified version of the Zuker's algorithm to compute in linear time the MFE hybridization of all positions to output multiple energetically favourable target sites. The advantage of this approach is it effectively reduces the problem to a local sequence alignment problem. With a time complexity proportional to $O(nm)$, where n and m are the length of the two sequences, this approach can effectively be applied to large sequence datasets, e.g. to find potential targets of a miRNA. However, the problem with this approach is it is biologically unreasonable to assume the interacting RNAs form only intermolecular base pairs. Intramolecular base pairs formed before interaction occurs could prevent intermolecular base pair formation.

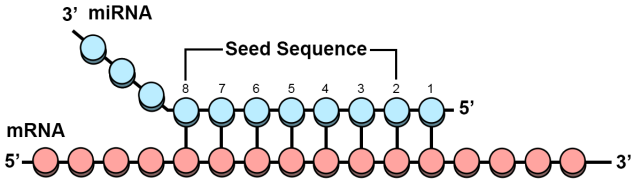

Fig. 3.8 *Seed sequence:* interaction of miRNA seed at position 2 - 8 with its target mRNA.

In the context of mRNA-microRNA interactions the *seed* sequence, figure 3.8, of the microRNA located at positions 2-8 from the 5' to 3' end is a very important feature for target recognition. There are several reported reasons why this is the case, the seed region is evolutionary conserved and identifies microRNA families that share common target mRNAs [106, 107]. Ameres *et al.* report stronger binding affinity of the seed region to the RNA induced silencing complex (RISC) [108]. In plants the microRNA seed sequence typically binds to its target to form a perfect match of Watson-Crick pairs, i.e. no gaps occur in the alignment and only `(GC)` and `(AU)` pairs occur. However, in animals it has been reported that gaps and `(GU)` wobble pairs can occur in the seed region [106, 109, 110]. Many variations of seed sequence have been reported, for example studies show shorter seed binding can correlate with target repression levels and an adenosine opposite position 1 of the microRNA as having a postive impact on site recognition [111]. Agarwal *et al.* rank perfect match Watson-Crick sites by decreasing conservation and efficacy [111].

MicroRNA target sites are most commonly located within the 3' UTR of mRNA, however there is experimental evidence of sites located in the coding and 5' UTR regions. One of the first microRNA target prediction tools `miRanda` [105] used to identify targets in *Drosophila* takes a three-step approach: (1) sequence complementarity is determined using a *Smith-Waterman* like alignment algorithm to score potential sites. The score is derived from seed rules that allow for gaps and wobble `(GU)` pairs. (2) The free energy of all identified sites are evaluated using the *Vienna RNA package*. (3) The final step is filtering of sites based on energy and optionally evolutionary conservation information.

An alternative approach to RNA-RNA interaction prediction takes into consideration both inter and intramolecular base pairs. This approach involves concatenating the two sequences and cofolding, using a modified Zuker's algorithm, to form a hybrid secondary structure. A major limitation of this approach is, for reasons similar to pseudoknot prediction, it is not possible to predict interactions on interior loops, such as hairpins and bulges, in the joint structure. Thus this approach excludes a large number of potential interaction sites [112]. An implementation of this approach with time complexity $O(n^3)$ can be found in the tool `RNAcofold` where in addition to the MFE structure McCaskill's algorithm is also incorporated. Thereby allowing calculation of base pair probabilities and calculation of the equilibrium centroid duplex structure [113]. Another tool taking this approach is `PairFold`, where Wuchty *et al.* algorithm for computing suboptimal secondary structures is incorporated into the prediction [114].

3.4.1 Accessibility based prediction

To overcome the limitations of the sequence only and cofolding approach, Muckstein *et al.* proposed in 2006 a two-step thermodynamic hybridization approach [112]: (1) removal of intramolecular base pairs from the target site, followed by (2) hybridization of the two molecules by formation of intermolecular base pairings:

$$\Delta G = \Delta G_{\text{open}} + \Delta G_{\text{duplex}} \tag{3.13}$$

Where ΔG_{open} is the energy contribution to remove intramolecular base pairings to open the target for binding. And, ΔG_{duplex} is the energy gained by

hybridization of the binding site. Muckstein *et al.* implement, in their tool `RNAup`, the accessibility based approach in the context of McCaskill's partition function [112]. Where for each subsequence interval $[i, j]$ the probability $P_u[i, j]$ that the iterval is unpaired is computed which is equivalent to ΔG_{open} and then for any valid interaction site the hybridization energy, ΔG_{duplex}, is calculated. For a maximum subsequence size w the asymptotic time complexity of their approach is $O(n^3 + nw^5)$ where n is the length of the sequence. In 2011 Bernhart *et al.* proposed an algorithm also based on McCaskill's partition function, `RNAplfold`, to compute accessibilities of all intervals in $O(n^3)$ time and $O(n^2)$ space [115].

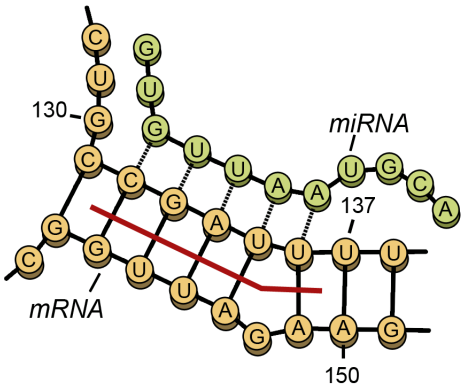

Fig. 3.9 *Target site accessibility: for interaction to occur at position 132 - 136 the intramolecular base pairs of the mRNA need to be destroyed.*

In 2007 Long *et al.* adapted the accessibility approach to predict microRNA targets and released the tool `STarMir` [116]. Their algorithm models interaction as a two-step process. Firstly, the microRNA forms base pairings with a block of four contiguous unpaired nucleotides, i.e. the nucleotides do not form intramolecular pairings. Then, the microRNA completes hybridization by disrupting any intramolecular base pairs. In their approach the authors statistically sample $1,000$ representative structures from the Boltzmann ensemble using `Sfold`.

In the same year `STarMir` was released Kertesz *et al.* investigated experimentally target site accessibility in the context of microRNA translational repression [117]. They show using *quantitative luciferase assay*, mutations affecting accessibility result in reduced microRNA action and conclude accessibility to be as important as seed match. They hypothesise genomes may have evolved such that target sites are positioned in unstructured UTR regions. Kertesz *et al.* incorporate their finding into a computational tool Probabil-

Tool	Main new feature	Ref.
TargetScan (2003)	A web-based tool that searches for miRNAs with conserved complementarity of 6mer, 7mer and 8mer seeds for a input mRNA.	[120]
miRanda (2003)	A dynamic programming algorithm to calculate free-energy of miRNA-mRNA duplexes. Intermolecular base pairs energy only, MicroRNA sequence is concatenated with the mRNA sequence at identified binding site.	[105]
RNAcofold (2006)	Introduces full free energy minimization algorithm for two input RNA sequences - intermolecular base pairs energy only.	[113]
STarMiR (2007)	Introduces the concept of target-site accessibility.	[116]

Table 3.1 Selection of popular microRNA target site prediction tools.

ity of Interaction by Target Accessibility (`PITA`). `PITA` firstly identifies seed match positions and then calculates for each position an accessibility score $\Delta\Delta G = \Delta G_{\text{duplex}} - \Delta G_{\text{open}}$. Finally, the statistical weight of all target site scores is calculated to give an overall interaction score.

Alkan *et al.* prove, by reduction from *the longest common subsequence of multiple binary strings*, the RNA interaction problem using the Nearest Neighbour Energy model to be NP-complete [118]. Importantly even when internal and external pseudoknot structures are excluded the problem is still NP-hard [118]. Recently, Lai *et al.* compared RNA-RNA interaction algorithms on an experimentally confirmed dataset and report the best performing tools to be energy based that take into consideration binding accessibility [119]. This result differs to structure prediction where comparative methods typically outperform. In their study the authors propose we may have reached the theoretical limits of a generalised RNA-RNA prediction algorithm. And suggest that further improvements may only come from algorithms designed for specific types of biological interactions.

MicroRNA target site prediction is difficult because exact mechanisms of target recognition are not fully understood. As stated by Cloonan, microRNA mediated repression of proteins by means of mRNA interaction at first appears to be an efficient regulatory control mechanism. However, biologically it would be more efficient to not produce the mRNAs in the first instance instead of producing and then destroying [121]. In summary, some of the challenges of microRNA target prediction are as follows:

- Experimentally detecting and validating microRNA interactions is a time-consuming and expensive tasks.

- Many types of seed matches, e.g. length, perfect matches, gaps and wobble pairs, etc. [106].
- Target site not always within the 3' UTR sequence. Considering full length mRNA transcripts is computationally challenging.
- microRNAs have a one to many relationship with their target mRNAs.
- As is the case with structure prediction interaction with other molecules, namely proteins such as the *RNA-induced silencing complex* RISC, are often ignored.

3.5 Conclusions

To summarise the most popular secondary structure prediction algorithm uses dynamic programming and parameterised energy model to predict a single minimum free energy structure. A number of variant structure prediction and interaction algorithms have been devised from this algorithm such as McCaskill's algorithm. Since development of McCaskill's partition function algorithm no novel single sequence prediction algorithm has been proposed that seriously challenges the energy-based framework in terms of accuracy and time complexity. Similiar to what has taken place at the sequence level, when the transition was made from Sanger-based sequencing to next-generation sequencing, a revolutionary new framework is needed to determine RNA secondary structures both computationally and experimentally. More recently, mainly since 2010 onwards, there has been growing interest in the development of models combining high-throughput structure probing data with computational prediction algorithms, see [122] for a detailed review on the subject. Although it has been reported incorporating this data can guide predictions to achieve better results it is not always the case [123]. The experimental techniques are noisy and currently have reproducability issues. Furthermore, they assume the native structure is the one of minimum free energy. As has been discussed throughout there is mounting evidence that RNAs have a more complex dynamic structural landscape. Therefore, current experimental approaches when successful will only provide a snapshot of structure for a particular moment in time for a specific environment. Likewise computational modelling of RNA folding presents many challenges and relies on progress in biology to better model nature's rules governing folding and interaction.

And, incorporating more rules into prediction tools is likely to lead to better accuracy but at the cost of computational complexity.

Chapter 4

Impact of SNPs on miRNA binding sites in metastable mRNAs

The contents of this chapter appear in the following publication:

[9] Luke Day, Ouala Abdelhadi Ep Souki, Andreas A. Albrecht and Kathleen Steinhöfel. "Accessibility of microRNA binding sites in metastable RNA secondary structures in the presence of SNPs", *Bioinformatics*, 30 (3): 343-352, 2014. doi: 10.1093/bioinformatics/btt695.

4.1 Introduction

In this chapter, our study on microRNA bindings to metastable RNA secondary structures close to minimum free energy conformations in the context of single nucleotide polymorphisms (SNPs) and messenger RNA concentration levels is presented. The aim of this study was to investigate whether features of microRNA bindings to metastable conformations could provide additional information supporting the differences in expression levels of two sequences defined by a SNP. We begin this chapter by reviewing important background literature related to our study. Firstly, we examine genetic variation caused by SNPs and their role in disease susceptibility (Sec. 4.2). In section 4.3 we discuss literature examining the crosslink between SNPs, microRNAs and disease by means of altered RNA structure and in section 4.4 we review literature on RNA expression levels. In section 4.5 we describe in detail the underlying hypothesis of our study. In section 4.6 our approach and dataset is presented and in section 4.7 we present our findings.

4.2 Genotypes, Phenotypes and Single Nucleotide Polymorphisms

Based on Gregor Mendel's theory of genetic inheritance all human cells, except for gametes or sex cells, are diploid meaning they contain two copies of each chromosome. This means all diploid cells contain a chromosome inherited from their father and a matching chromosome from their mother, resulting in homologous chromosomes. Therefore, each cell will have two copies of each gene. However, the two genes may not be identical. For example, consider figure 4.1 showing a pair of homologous chromosomes.

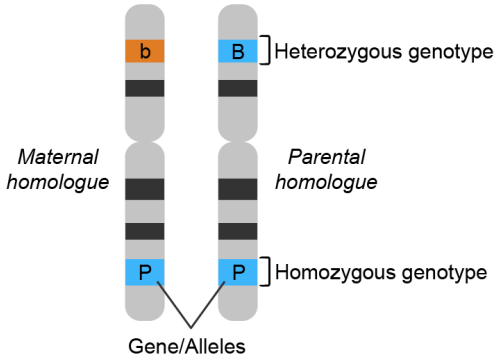

Fig. 4.1 Homologous chromosomes: when two alleles do not match between a pair of chromosomes they form a heterozygous genotype. And, when alleles match they form homozygous genotype.

Assume the boxes labelled b and B are the MC1R gene that is responsible for red hair colour. The genetic code of the MC1R gene must have different variations to signify red hair and no red hair. Variations of a gene are called alleles. Alleles make up an individual's genotype that results in some observable trait referred to as a phenotype, e.g. red hair. Every human has two copies of the MC1R gene, one from each parent, and the gene can have two variations. Namely, red and not red. Therefore, there are four possible combinations of the MC1R gene. If b represents red and B represent not red, then there can be bb, bB, Bb and BB. When a chromosome has two differing variations of a gene it is said to be heterozygous. In diploid organisms, both copies of a gene are typically expressed simultaneously, known as biallelic expression. And, a small number of genes only one allele is expressed, known as monoallelic expression. Biallelic expressed genes, such as such MC1R, one allele typically dominates the other allele. For example, in the bB case

the B (not red) allele dominates the b (red) allele. Meaning, the phenotype of the individual represented by the example would not have red hair. In this particular example, a person would only have red hair if both alleles are bb. A unique genotype is formed by genetic variations such as copy number variants (CNVs), indels (deletions or insertions), Single Nucleotide Polymorphisms (SNPs) and structural variants resulting from meiosis. Genetic recombination of both sets of chromosomes ensure genetic information from both the parental and maternal lineage is inherited by offspring.

4.2.1 Single Nucleotide Polymorphisms

The construction of new alleles caused by genetic variations is the driving force of biological change or evolution. Single nucleotide polymorphisms (SNPs) are single mutations of DNA nucleotides that occurred sometime in evolutionary history and now found with high frequency in the genomes of a population.

Fig. 4.2 Single Nucleotide Polymorphisms: A single nucleotide mutation. In this example, the A-allele is the wild-type because it occurs more frequently than the G-allele. Assumption, is at some point in evolution the nucleotide adenine mutated to a guanine $(\mathrm{A} \rightarrow \mathrm{G})$.

A set of alleles or is known a haplotype. The most frequently observed allele of a population is referred to as the wild-type or major allele. And, any less frequent variants are typically referred to as minor-alleles or simply variant alleles. At the DNA sequence level human genomes on average differ by approximately 2 - 4 million nucleotides making any two humans about 99.9% identical [124]. SNPs make up a major proportion of the remaining 0.1%. SNPs are the most common type of genetic variation occurring approximately once every 100 - 300 nucleotides [125]. Thus far over 100 million human SNPs have been identified and validated (dbSNP Build 147) [126]. Genetically SNPs are what make us all unique and play a role in determining phenotypic traits such as height, skin, hair and eye colour, personality and blood type.

4.2.2 SNPs and disease

Reich *et al.* and references within discuss the common disease/common variant hypothesis, i.e. the idea that common diseases are likely to have common genetic variations in a population [127]. During the last decade this hypothesis has been put to the test. The Human Haplotype Map project (HapMap) launched in 2003 mapped allele frequencies and correlation patterns among variants, allowing for examination of linkage disequilibrium. Linkage disequilibrium is a property of SNPs on a contiguous stretch of DNA that describes the degree to which an allele defined by a SNP is correlated with an allele defined by another SNP with a population [128]. A similar project, the 1000 Genomes Project Consortium set up in 2008 investigated the degree of genetic variation in the human population. Recently the project sequenced the genomes of 2,504 people across five continental regions to give the most comprehensive global picture of genetic variation to date. They recently reported finding over 88 million variants, including 84.7 million single nucleotide polymorphisms of which >99% have frequency >1% [129]. The 1000 genomes consortium report a typical genome differs from the reference human genome at 4.1 to 5.0 million sites.

Growth in genomic variant data has led to a large number of Genome Wide Association Studies (GWAS). The ultimate aim of these studies are to identify and map genetic variant risk factors to specific diseases. With the hope of using this information to predict who is more susceptible to develop disease. Over the past few years several new SNP-trait associated genes have been identified. For example, the National Human Genome Research Institute (NHGRI) GWAS catalog reports 489 SNP-trait associations for cancer and 169 for cardiovascular disease [130]. SNPs are also a common cause of monogenic genetic diseases such as cystic fibrosis, Parkinson's disease and sickle cell anaemia. Sickle cell anaemia is a disease caused by abnormal haemoglobin protein. Sickle cell anaemia is known to be caused by a SNP occurring in the beta-globin (HBB) gene, a subunit of haemoglobin. The SNP $(\mathrm{A} \rightarrow \mathrm{T})$ results in a change to an amino acid when HBB mRNA is translated into a protein. The single change of amino acid results in red blood cells becoming sickle shaped and unable to flow efficiently around the body. This disease is most prevalent within ethnic groups having ancestry from hot climates.

For genotype (A; A) no disease occurs and all red blood cells are normal shaped. Under normal conditions no disease is also observed with the carrier

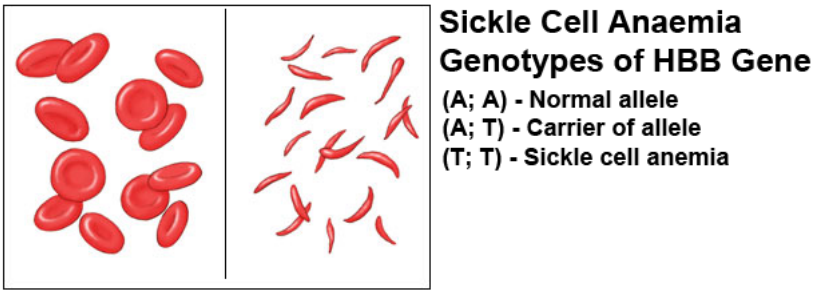

Fig. 4.3 (Left) the structure of normal red blood cells. (Right) Sickle shaped red blood cells. If both the maternal and paternal HBB gene contain the T base SNP then the individual will develop sickle cell anaemia.

heterozygous genotype (A; T), or sickle cell trait, because one allele is normal. People with the sickle cell trait will however have some sickle shaped red blood cells. If both alleles contain the SNP then expression of either gene results in a dysfunctional protein. A proposed reason for this disease associated SNP is it may provides an evolutionary survival advantage against malaria. A hypothesis first proposed in 1954 by A. C. Allison who observed significantly lower malaria infection rates in children with the sickle-cell trait in Africa [131]. Numerous clinical studies support the malaria resistance hypothesis, however the underlying mechanism of how this is achieved remains unclear [132, 133]. Sickle cell anaemia is clear example of the relationship between genomic sequence, structure, function and disease. It is evident from the growing number of GWAS studies that genetic variants, namely SNPs, play an important in common disease susceptibility. And taking into consideration recent advances in non-protein coding genomics and decreasing sequencing costs these studies highlight the importance of personalised medicine, i.e. treatment based on one's own genotype. A major problem with GWAS is elucidating the underlying mechanisms of how these genes increase risk.

4.3 miR-SNPs and RNA secondary structure

One area of research into a potential mechanism on how SNPs influence disease is miRNA-SNPs [134]. miRNA-SNPs occur either within a microRNA gene or target mRNA gene. Over recent years several studies have proved miRNA-SNPs can lead to a loss of gene regulation. One of many such studies is Chin *et al.* who proved a SNP in the 3' UTR of the KRAS oncogene changes

binding of the tumor suppressor microRNA let-7; resulting in increased risk of lung cancer [135].

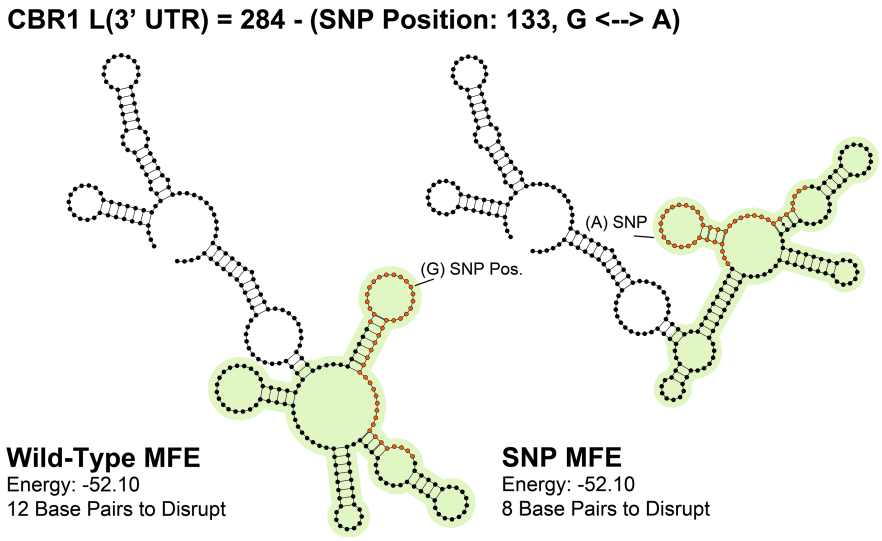

Fig. 4.4 Predicted MFE conformations for two variant 3' UTR transcripts of gene CBR1.

A suggested explanation for the crosslink between SNPs, microRNAs and disease is that SNPs modify microRNA binding sites by altering structure. It has long been known that SNPs impact mRNA structure [136]. This has led to efforts to develop algorithms to characterise potential SNP effects on RNA structure. Churkin *et al.* proposed the tool `RNAmute` which calculates all single point mutations for a given input RNA sequence, of length in order of 100 - 150 nt, and organises them according to similiarity to the wild type structure [137]. Waldispühl *et al.* propose, by means of a modified version of the McCaskill algorithm, a tool `RNAmutants` to probe the mutational landscape [138]. `RNAmutants` calculates for a given RNA sequence of length n and integer $k_{\max} \leq n$ ensembles for the k-neighbourhood, i.e. the partition Z_k over all secondary structures of all $k-point$ mutatants in $O(n^5)$ time and $O(n^3)$ space. Halvorsen *et al.* proposed in 2010 the algorithm `SNPfold` [139] to calculate paring probabilities using a partition function for wild and SNP variant sequences to study disease associated SNPs. `SNPfold` uses the Pearson correlation of pairing probabilities to identify "RiboSNitches", SNPs having significant effect on structure. More recently, Sabarinathan *et al.* proposed the algorithm `RNAsnp` to calculate local secondary structure changes induced by a SNP [140]. `RNAsnp` uses precomputed tables of SNP effects and incorporates `RNAplfold` [115] to compute effects on large sequences and datasets.

Nicoloso *et al.* hypothesised in 2010, SNPs occurring within microRNA target sites can influence tumour susceptibility. In their study, they analysed *in silico* and *in vitro* SNPs associated with breast cancer risk. Using the HapMap dataset (version 21) and microRNA prediction tool miRanda the author's conducted a genome-wide analysis of transcribed SNPs. And, predict that 64% of SNPs can modify the binding energy of putative microRNA-mRNA interactions by over 90% [141]. Mallick *et al.* investigated using a bioinformatics approach the impact of SNPs in microRNA binding sites on genes related to Alzheimer's disease prognosis [142]. The authors analysed the interaction of 166 microRNAs experimentally reported to be differentially expressed, i.e. up and down regulated, in brain tissues of Alzheimer's disease patients. And, report allelic variant mRNAs defined by a SNP result in stronger or weaker microRNA-mRNA binding over MFE structures. The authors hypothesise SNPs ability to modify, create or destroy microRNA binding sites may contribute to Alzheimer's disease pathogenesis. Haas *et al.* investigated experimentally and computationally the impact of SNPs in the 3' UTR of mRNAs *angiotensin II receptor type 1* (AGTR1; A $\leftrightarrow$ C) and *Muscle RAS Oncogene Homolog* (MRAS; C $\leftrightarrow$ T) in the context of microRNA binding [143]. Haas *et al.* hypothesise 3' UTR miR-SNPs play an important role microRNA-mediated gene regulation by means of structural alteration resulting in change to microRNA binding site accessibility.

The impact of SNPs on minimum free energy mRNA conformations has been comprehensively studied in Johnson *et al.* [144]. The authors analysed a total number of 34,557 SNPs in 12,450 genes. The minimum free energy conformations were calculated by using `RNAfold`. The authors provide a great variety of data about the distribution of SNPs within mRNA transcripts and their effect on minimum free energy values of secondary structures as well as on the profile of the ensemble of suboptimal structures and structures with high Boltzmann probabilities. The authors compare structures between major and minor alleles for biallelic SNPs and report the majority alter MFE and suboptimal structures; with 34.1% of minor alleles having MFE structures identical to the major allele and only 6.4% near identical structure ensembles. The authors analysed further the 22,785 (65.9%) SNPs with predicted structural change and found the greatest frequency of structural change comes from transversions involving guanine, i.e. (G $\leftrightarrow$ U) and (G $\leftrightarrow$ C). Martin *et al.* study structural changes induced by SNPs in the 5' UTR of the human FTL (Ferritin Light Chain) gene in conjunction with associated SNPs that restore

the overall wild-type ensemble of secondary structures, thus leading to the notion of structure-stabilising haplotypes [145]. The authors also analysed the stabilising effect of multiple structure-stabilising haplotypes on binding sites of microRNAs and RNA binding proteins (nine cases of 3' UTRs and one case of 5' UTR). As pointed out by the authors, the findings suggest that certain SNP pairs are conserved in the human population because they stabilise ensembles of mRNA conformations.

4.4 RNA expression levels

Within the past few years, analyzing concentration levels of microRNAs (miRNAs) and their putative messenger RNA (mRNA) targets has become a major topic in miRNA research. Subkhankulova *et al.* experimentally evaluated a parameterised analytical expression that estimates the number of genes g having t transcripts present in a single cell [146]. The parameters are adjusted based on microarray data for a large number of genes extracted from single embryonic mouse neural stem cells, where the actual aim is the comparison of transcript numbers in phenotypically identical cells. The authors conclude from observed data for about 13,000 genes that the typical number of gene copies lies between 5 and 20, with 85% of genes having less than 100 copies in a single cell. Although the analysis is carried out for a specific cell type, the authors expect similar distribution results for a wide range of cell types. Arvey *et al.* provide experimental evidence that short RNAs (microRNAs and siRNAs) having a higher number of target transcripts within a single cell will downregulate each individual target gene to a lesser extent than those with a lower number of targets, which implies that the competition between target genes for a limited number of small RNAs may determine the degree of downregulation [147]; see also Salmena *et al.* for the concept of competing mRNAs associated with the number and distribution of multiple binding sites [148].

Salmena *et al.* provide data supporting the assumption that endogenous miRNAs preferentially target mRNAs with very long 3' UTRs [148]. The authors also discuss - among other features - the interaction between exogenous and endogenous miRNAs and critically assess the value of microarray data in the context of miRNA target prediction. The sequencing method developed by [149] supports the assertion that only the most abundant miRNAs

mediate target suppression. For example, deep sequencing of monocyte cells revealed the presence of about 310 miRNAs, with only about 40% of the miRNAs showing suppressing activity. For more than 80% of the targets, the corresponding miRNA was expressed above 100 reads per million (RPM). For the miRNA target prediction tool `TargetScan`, Garcia *et al.* demonstrate how predictions may improve if target abundance is accounted for in binding scores [150]. The impact of the life cycle of mRNAs on siRNA and microRNA efficacy is studied in [151]. The authors draw the conclusion that microRNA target prediction could be improved if data about mRNA turnover rates are incorporated into prediction tools. While [147, 151] focus on mRNA concentration levels, [152] propose different parameterised models of RNA interference that describe the effects of varying quantities of siRNAs. The models are derived from the basic equation $\mathrm{d}X_m/\mathrm{d}t = k_m - d_m X_m - \delta(X_m, X_s)$, where X_m and X_s are the mRNA and siRNA concentrations, k_m is the mRNA transcription rate, d_m the basal mRNA degradation rate, and $\delta(X_m, X_s)$ is the siRNA induced mRNA degradation rate. The models differ in the assumption about $\delta(X_m, X_s)$ and are fitted to experimental data obtained for a single siRNA targeting the coding region of the EGFP mRNA. The authors obtain the best fit for $\delta(X_m, X_s) = pX_m X_s^h/(q^h + X_s^h)$ with $h \sim 4.5$, $p \sim 0.008/\mathrm{min}$, and $q \sim 0.1\mathrm{pmol}$. Within a similar framework, Osella *et al.*[153] study the so-called miRNA-mediated feedforward loop in which a master transcription factor regulates a miRNA together with a set of target genes, and the mathematical models studied by Loinger *et al.* [154] additionally account for the concentration level of the argonaute protein complex. Baker *et al.* [155] study analytically the impact of multiple small non-coding RNAs on the regulation of a single target mRNA and subsequently the dynamics of protein production. Ragan *et al.* combine the concept of miRNA binding site accessibility with miRNA and mRNA concentration levels [156]. For $[S]$, $[T]$ and $[ST]$ denoting the equilibrium (final) concentrations of the miRNA, target mRNA and of the hybridized structure, respectively, the authors utilise the equilibrium condition $[ST]/([S][T]) = exp(-\Delta\Delta G/c)$, where $\Delta\Delta G$ is the energy score that accounts for making the binding site accessible and the free energy of the hybridized structure (the constant c stands for the product of the gas constant and the temperature). Combining the equilibrium condition with conservation of mass equations (for initial concentrations $[S_0]$ and $[T_0]$) eventually leads to an analytical expression for $[S]$ in terms of $[S_0]$, $[T_0]$ and $\Delta\Delta G$, where the latter is calculated for a particular binding site.

Marin and Vaniŏek introduce a new accessibility-based algorithm that utilises a statistical analysis of all putative binding sites for a given miRNA-3' UTR pair [157]. Among the top 100 target predictions for 153 fruit fly miRNAs, the algorithm finds more than twice as many validated targets compared to other accessibility-based target prediction methods. Reviews of existing miRNA target prediction tools and information about latest developments can be found in [158] and [159]. The target prediction tool `CoMeTa` designed by [160] operates on the assumption that targets of a given miRNA are co-expressed with each other. The target prediction score is based upon the evaluation of thousands of publicly available microarray data. For the 675 human miRNAs analysed in the study, more than 90% of the validated targets fall within the first 50% of predicted targets (which, however, could be a large number). In a similar way, the tool `miRror` designed by [161] combines scores produced by an ensemble of established miRNA target prediction tools with rankings obtained from gene expression and HITS-CLIP data. Johnson *et al.* consider the problem of selecting an antisense sequence that is able to effectively bind to a target mRNA and block protein synthesis [162]. One of the key features of the authors' method is the presuppositions that mRNA secondary structures are in a constant state of flux and are assuming different suboptimal states. Johnson *et al.* designed a tool that generates and compares suboptimal secondary structures of a given mRNA sequence [162] . The comparison aims at identifying regions that are least 'similar' among the set of folded structures, which indicates volatility in intramolecular hydrogen bonding. Such regions are seen as candidates for antisense binding. The method is evaluated on six mRNA sequences and compared to results produced by the `Soligo` application of `Sfold` [163].

4.5 Hypothesis

SNP occurring within close proximity of a miRNAs target site could disrupt the duplex binding energy or make the target site inaccessible for microRNA interaction by means of local structure change. Previous computational analyses on the impact of SNPs on microRNA binding site accessibility, such as those reported by Haas *et al.* [143] and Mallick *et al.* [142] described earlier, consider only the structural impact of SNPs on MFE conformations. If RNA folds into an ensemble of low energy states close to the MFE, as evidence

suggests [85, 162], then in the context of microRNA binding is the SNP allele showing increased/decreased expression more/less accessible for microRNA binding over metastable conformations?

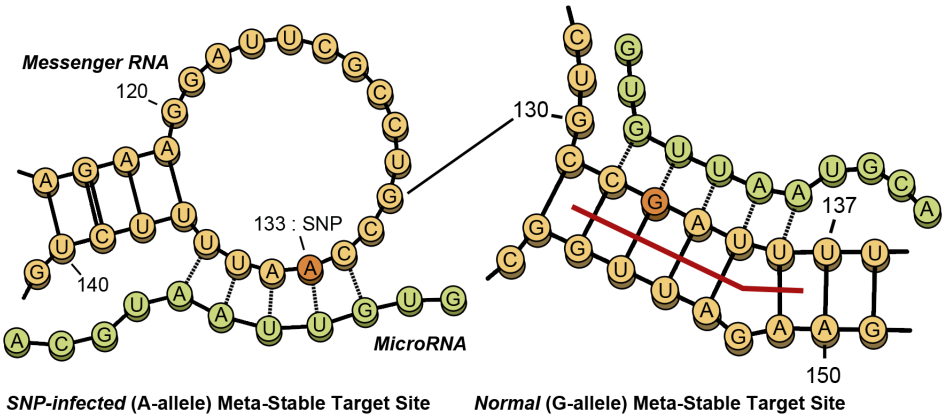

Fig. 4.5 MicroRNA binding site accessibility.

Figure 4.5 illustrates how a SNP occurring within a microRNA target site could change allelic mRNA expression. On the left is a SNP-infected metastable structure with with adenine at position 133 and microRNA binding site occurring at a hairpin and on the right is the wild-type allele with guanine at position 133 and the target site occurring at a helix. The SNP-infected allele is in a more accessible state for microRNA binding. It could be the case that the MFE structure of the wild-type allele has a more accessible target site and stronger predicted binding than the SNP infected allele. However, it is also possible that the SNP-infected allele has a greater number of metastable structures with high energy barriers and more accessible target sites than the wild-type allele.

In our study we investigated microRNA bindings to sets of metastable secondary structures induced by 3' UTRs and their mutated counterparts, where the single nucleotide polymorphisms (SNPs) are located within the microRNA binding site. The research question motivating this study is the following: do metastable structures of the allele reported to have stronger inhibitory effect have more accessible microRNA binding sites? i.e. does a SNP within a microRNA binding site alter microRNAs ability to regulate expression over metastable conformations? One of the main motivations for this study is the possibility to improve microRNA-mRNA predictions by incorporating mRNA copy numbers and metastable target site accessibility information into the algorithm.

4.6 Approach

Sethupathy and Collins critically assess several genetic association studies related microRNA bindings and the potential impact of SNPs in binding regions [134]. The authors highlight the importance of follow-up functional experiments for a deeper understanding of the real effect microRNA target site variations may have on the development of various human diseases. Consequently, we researched for our analysis recent literature for studies on the impact of SNPs on microRNA bindings meeting the following four conditions:

1. the expression levels of both alleles involved are analysed experimentally by SNP genotyping (for some instances, a combination of clinical association studies and strong *in silico* results) or related methods involving PCR experiments,

2. the underlying allele information can be extracted as consistent data from the NCBI database, the dbSNP database and the Ensembl database (3' UTR transcripts),

3. the microRNA bindings are predicted by the latest version of `STarMir` at least for the allele with the stronger inhibitory effect with the SNP position being inside the binding region and

4. identifying all metastable conformations within an energy offset E above MFE conformations is computationally feasible (3' UTR length $\leq$ 1,000 nt).

Meeting all four conditions reduced the number of case studies we found to a small set of 14 [mRNA; SNP; miRNA] cases of 3' UTR lengths ranging from 124 up to 1078 nt. In our study, the instances [mRNA/3' UTR; SNP; miRNA] were selected based on strong expression level analyses, SNP locations within binding regions and the computationally feasible identification of metastable conformations. The cases are summarised in Table 4.1 on page 61. The data in Table 4.1 are sourced from the following publications:

1) **[LIG3; rs4796030; miR-221]**: The LIG3 gene is involved in DNA strandbreak repair pathways and is analysed by Teo *et al.* [164]. The SNP rs4796030 is defined by A $\leftrightarrow$ C at 3' UTR position 83 of NM_002311.4. The 3' UTR length is 124 nt. The authors studied bladder cancer cases

by using clinical association studies combined with *in silico* target predictions. They conjecture a stronger inhibitory effect of miR-221 for the C-allele.

2) **[CBR1; rs9024; miR-574-5p]**: The CBR1 gene encodes an enzyme that catalyzes a wide variety of carbonyl compounds. The case is analysed by Kalabus *et al.* [165]. The SNP rs9024 is defined by G $\leftrightarrow$ A at 3' UTR position 133 of NM_001757.2. The 3' UTR length is 284 nt. The authors employed dual-luciferase assays to evaluate the miRNA binding to both alleles and observed a stronger inhibitory effect of miR-574-5p for the A-allele.

3) **[HTR3E; rs56109847; miR-510-5p]**: This case is analysed by Kapeller *et al.* in [166], see also Sethupathy and Collins [134]. The SNP rs56109847 (rs62625044) is defined by G $\leftrightarrow$ A at 3' UTR position 76 of NM_001256614.1. The 3' UTR length is 302 nt. The authors measure luciferase activity to evaluate the miRNA binding to both alleles and observe a stronger inhibitory effect of miR-510-5p for the G-allele, see Figure 1B in [166].

4) **[HLA-G; rs1063320; miR-148a-3p]**: This case is analysed by Tan *et al.* in [167]. The SNP rs1063320 is defined by C $\leftrightarrow$ G at 3' UTR position 233 of NM_002127.5. The 3' UTR length is 386 nt. The study aims at exploring factors affecting the asthma risk. The authors used real-time PCR and luciferase assays for measuring expression levels of miRNAs and C/G-alleles and found evidence for a stronger inhibitory effect of miR-148a-3p for the G-allele.

5) **[PARP1; rs8679; miR-145-5p]**: This case is analysed by Teo *et al.* in [164]. The SNP rs8679 is defined by T $\leftrightarrow$ C at 3' UTR position 607 of NM_001618.3. The 3' UTR length is 769 nt. The methodology is the same as for [LIG3; rs4796030; miR-221], and the authors presume an additive effect of both instances on bladder cancer risk. For PARP1/rs8679, the authors conjecture a stronger inhibitory effect of miR-145-5p for the T-allele.

6) **[WFS1; rs1046322; miR-668-3p]**: This case is analysed by Kovacs-Nagy *et al.* in [168]. The SNP rs1046322 is defined by G $\leftrightarrow$ A at 3' UTR position 253 of NM_001145853.1. 3' UTR length is 779 nt. Expression

levels of luciferase assays are measured for both alleles, with a stronger inhibitory effect of miR-668-3p for the G-allele.

7) **[IL-23R; rs10889677; let-7e]**: This case is analysed by Zwiers *et al.* in [169]. The SNP rs10889677 is defined by C ↔ A at 3' UTR position 309 of NM_144701.2. The 3' UTR length is 851 nt. The authors study risk factors for inflammatory bowel diseases. Wang *et al.* associate [IL-23R; rs10889677] with breast cancer development [170]. Zwiers *et al.* employ real-time PCR and luciferase assays for measuring expression levels, and the authors conclude a stronger inhibitory effect of let-7e for the C-allele [169].

8) **[RYR3; rs1044129; miR-367]**: This case is analysed by Zhang *et al.* in [171]. The SNP rs1044129 is defined by A ↔ G at 3' UTR position 839 of NM_001036.3. The 3' UTR length is 880 nt. The authors study risk factors of breast cancer development. The authors employ real-time PCR and luciferase assays for measuring expression levels. The authors observe a stronger inhibitory effect of miR-367 for the A-allele.

9) **[AGTR1; rs5186; miR-155-5p]**: This case is analysed by Haas *et al.* in [143]. The SNP rs5186 is defined by A ↔ C at 3' UTR position 86 of NM_032049.3. The 3' UTR length is 888 nt. The authors analyse luciferase assays for measuring expression levels and observe a stronger inhibitory effect of miR-155-5p for the A-allele, see 'long case' in Figure 2C in [143].

10) **[FGF20; rs12720208; miR-433-3p]**: This case is analysed by Wang *et al.* in [172]. The SNP rs12720208 is defined by C ↔ T at 3' UTR position 182 of NM_019851.2. The 3' UTR length is 903 nt. The authors analyse luciferase assays for measuring expression levels and observe a stronger inhibitory effect of miR-433-3p for the C-allele. Note that in this specific case we used the submission ss20399075 instead of the default dbSNP entry ss28476621 for consistency with the NCBI entry of NM_019851.2 and the corresponding 3' UTR transcript entry ENST00000180166 at ENSEMBL.

11) **[HOXB5; rs9299; miR-7-5p]**: This case is analysed by Luo *et al.* in [173]. The SNP rs9299 is defined by G ↔ A at 3' UTR position 141 of NM_002147.3. The 3' UTR length is 952 nt. The miRNA-3' UTR

binding is studied in the context of bladder cancer development. Both real-time PCR and luciferase reporter assays are applied gene expression measurements. The authors observe a stronger inhibitory effect of miR-7-5p for the A-allele.

12) **[RAD51; rs7180135; miR-197-3p]**: This case is analysed by Teo *et al.* in [164]. The SNP rs7180135 is defined by G $\leftrightarrow$ A at 3' UTR position 718 of NM_002875.4. The 3' UTR length is 978 nt. The methodology is the same as for [LIG3; rs4796030; miR-221] and [PARP1; rs8679; miR-145-5p]. The authors conjecture a stronger inhibitory effect of miR-197-3p for the G-allele.

13) **[ORAI1; rs76753792; miR-519a-3p]**: This case is analysed by Chang *et al.* in [174]. The SNP rs76753792 is defined by C $\leftrightarrow$ T at 3' UTR position 86 of NM_032790.3. The 3' UTR length is 1034 nt. The authors study the susceptibility of atopic dermatitis in Japanese and Taiwanese populations. Among other methods, real-time PCR is applied to gene expression analysis. Chang *et al.* mention the impact of miRNAs as subject of future research, i.e. no specific miRNA is identified. Based upon miRNA target predictions for ORAI1 and the SNP position we selected miR-519a-3p for the present study. For miR-519a-3p, the binding prediction returned by `StarMir` is stronger for the C-allele.

14) **[RAP1A;rs6573;miR-196a]**: This case is analysed by Wang *et al.* in [175]. The SNP rs6573 is defined by A $\leftrightarrow$ C at 3' UTR position 366 of NM_002884.2. The 3' UTR length is 1078 nt. The authors study how rs6573 affects the risk of esophageal squamous cell carcinoma. The regulatory function of miR-196a is analysed by luciferase reporter assays. The authors conclude a stronger inhibitory effect of miR-196a for the A-allele.

See Table 4.1 for an overview of stronger and weaker alleles reported by the above studies.

No.	Symbol	Gene	NCBI Ref.	L(3' UTR)	w-allele	s-allele	dbSNP ID	ENSEMBL ID
1.	LIG3	DNA Ligase 3	NM_002311.4	124	A	C	rs4796030	ENSG00000005156 ENST00000262327
2.	CBR1	Carbonyl Reductase 1	NM_001757.2	284	G	A	rs9024	ENSG00000159228 ENST00000290349
3.	HTR3E	5-Hydroxytryptamine Receptor 3E	NM_001256614.1	302	A	G	rs56109847	ENSG00000186038 ENST00000335304
4.	HLA_G	Major Histocompatibility Complex	NM_002127.5	386	C	G	rs1063320	ENSG00000204632 ENST00000360323
5.	PARP1	Poly(ADP-Ribose) Polymerase 1	NM_001618.3	769	C	T	rs8679	ENSG00000143799 ENST00000366794
6.	WFS1	Wolframin ER Transmembrane Glycoprotein	NM_001145853.1	779	A	G	rs1046322	ENSG00000109501 ENST00000226760
7.	IL-23R	Interleukin 23 Receptor	NM_144701.2	851	A	C	rs10889677	ENSG00000162594 ENST00000347310
8.	RYR3	Ryanodine Receptor 3	NM_001036.3	880	G	A	rs1044129	ENSG00000198838 ENST00000389232
9.	AGTR1	Angiotensin II Receptor Type 1	NM_032049.3	888	C	A	rs5186	ENSG00000144891 ENST00000497524
10.	FGF20	Fibroblast Growth Factor 20	NM_019851.2	903	T	C	rs12720208	ENSG00000078579 ENST00000180166
11.	HOXB5	Homeobox B5	NM_002147.3	952	G	A	rs9299	ENSG00000120075 ENST00000239151
12.	RAD51	RAD51 Recombinase	NM_002875.4	978	A	G	rs7180135	ENSG00000051180 ENST00000267868
13.	ORAI1	Release-Activated Calcium Modulator 1	NM_032790.3	1034	T	C	rs76753792	ENSG00000182500 ENST00000330079
14.	RAP1A	Member of RAS Oncogene Family	NM_002884.2	1078	C	A	rs6573	ENSG00000116473 ENST00000369709

Table 4.1 Reference numbers from NCBI database (mRNA wild-type), ENSEMBL and dbSNP database.

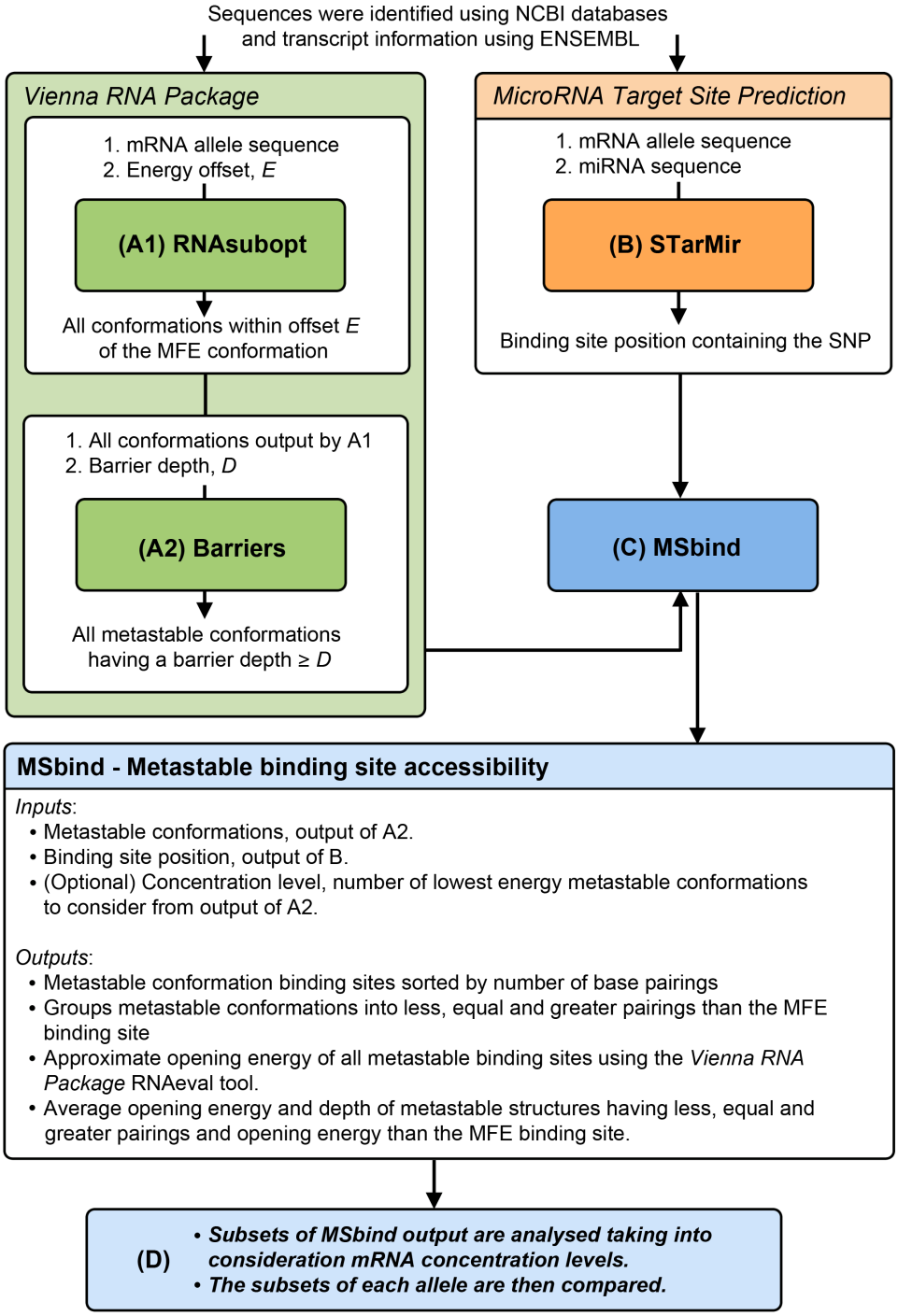

Fig. 4.6 Flow of analysis.

In our study we assume the existence of multiple active RNA conformations, as reported by Solomatin *et al.* [85] and Johnson *et al.* [162] and literature therein, instead of a single MFE conformation as the biologically active state. The second basic feature of our approach relates to the presence of multiple copies of each individual mRNA. In more detail, we proceed as follows for a given [mRNA; SNP; miRNA] instance:

(**A**) For both alleles (3' UTR and RS), the sets MS(3' UTR, δE) and MS(RS, δE) of meta-stable states within an energy offset δE above the mfe-conformation are identified by using `RNAsubopt` and `Barriers` [65].

(**B**) `STarMir` [116] is applied to both alleles for the given miRNA. Although the 14 basic cases [mRNA;RS;miRNA] selected from recent literature were analysed in the corresponding publications by prediction tools different from `STarMir`, we obtained for each of the cases at least for one allele a binding site predicted by `STarMir` with the SNP position inside.

(**C**) For the predicted binding site BS(miRNA), the elements of the sets MS(3' UTR, δE) and MS(RS, δE) are examined with respect to certain features, such as the number of base pair bindings within BS(miRNA) and the approximate free energy ΔG_{BS} of bindings within BS(miRNA) according to standard data of the Nearest Neighbour Model.

(**D**) Subsets of MS($\cdots, \delta$E) are analysed in the context of the number of mRNA copies, as analysed in [146], in the same way as in (C), i.e. for $k = 20, 60, 100$ the sets MS($\cdots, k$) $\subset$ MS($\cdots, \delta$E) are examined, where k indicates the assumption about the number of mRNA copies. Note that the number of copies is different from concentration levels, which are measured, e.g., *per mol*, see [152].

A flow chart of the approach is given on page 62.

4.7 Results

When executing step (A), we applied the standard settings of `RNAsubopt` and `Barriers`, which includes that isolated base pairs are not admitted in secondary structures and free energy values are discriminated with an accuracy of 0.1 kcal/mol. The setting of δE depends on the length of the 3' UTR and

was selected in such a way that a sufficiently large number of meta-stable conformations is available for analysing $\mathrm{MS}(\cdots,k)$ with $k \leq 100$. Along with the energy offset δE, we tried to restrict meta-stable states to 'deep' local minima: The parameter D indicates the 'depth' of a local minimum or 'escape height' from a local minimum, which is taken from the barrier tree as the distance to the nearest saddle point. By $|\mathrm{SecStruc}_{w/s}|$ (short for both cases of $|\mathrm{SecStruc}_{weak}|$ and $|\mathrm{SecStruc}_{strong}|$) we denote the number of secondary structures returned by `RNAsubopt` for an offset δE above the mfe-conformation, where the index w indicates the allele with the weaker and s with the stronger miRNA inhibitory effect (binding prediction - for [ORAI1; rs76753792; miR-519a-3p]). Analogously, $|\mathrm{MS}_{w/s}|$ is the number of local minima within the δE range with an 'escape height' larger or equal to D.

4.7.1 Number of metastable conformations

The results of step (A) are summarised in Table 4.2. For nine out of the fourteen instances, the values of $|\mathrm{MS}_w|$ and $|\mathrm{MS}_s|$ are relatively close or (much) larger for $|\mathrm{MS}_s|$, see Figure 4.7. For the remaining instances, the ratio $|\mathrm{MS}_w|/|\mathrm{MS}_s|$ ranges from 1.31 (FGF20) up until 6.47 (AGTR1). Thus, the number of secondary structures classified as meta-stable conformations *per se* does not discriminate between the two cases of weaker and stronger binding to the associated miRNA.

The correlation between $|\mathrm{MS}_w|$ and $|\mathrm{MS}_s|$ does not necessarily extend to $|\mathrm{SecStruc}_{w/s}|$. For example, for LIG3 we have $|\mathrm{MS}_w|/|\mathrm{MS}_s| > 1$, whereas $|\mathrm{SecStruc}_w|/|\mathrm{SecStruc}_s| < 1$.

4.7.2 MicroRNA binding sites and energy predictions

The results obtained in step (B) are summarised in Table 4.3. For a given input [3' UTR/SNP; miRNA], `STarMir` returns a large number of data items and graphical representations of miRNA binding patterns. We focus on the binding regions and four energy values:

(a) ΔG_{nucl} relates to the assumption that the initial stage of base-pairing (nucleation) requires a gain in free energy that is greater than the energy cost for the translational and rotational entropy loss when both miRNA and mRNA are fixed in a conformation by intermolecular base-pairing. The value of ΔG_{nucl} is calculated by using a sample of 1000 structures

	LIG3	CBR1	HTR3E	HLA-G	PARP1	WFS1	IL-23R
L(3' UTR)nt	124	284	302	386	769	779	851
w-allele	A	G	A	C	C	A	A
s-allele	C	A	G	G	T	G	C
δE kcal/mol	6.0	6.0	6.0	4.0	3.0	2.7	2.0
D kcal/mol	1.2	1.4	1.4	1.4	1.2	0.8	1.2
$\lvert\text{SecStruc}_w\rvert$	2.1×10^4	1.1×10^7	2.4×10^5	9.7×10^6	1.6×10^6	1.1×10^7	7.6×10^5
$\lvert\text{SecStruc}_s\rvert$	2.9×10^4	1.6×10^7	2.4×10^5	8.4×10^6	1.1×10^7	1.1×10^7	4.0×10^5
$\lvert\text{MS}_w\rvert$	349	7,457	996	11,957	1,709	79,273	1,080
$\lvert\text{MS}_s\rvert$	317	11,187	1,173	10,473	14,281	79,577	964

	RYR3	AGTR1	FGF20	HOXB5	RAD51	ORAI1	RAP1A
L(3' UTR)nt	880	888	903	952	978	1034	1078
w-allele	G	C	T	G	A	T	C
s-allele	A	A	C	A	G	C	A
δE kcal/mol	2.3	2.3	2.3	2.0	2.2	2.0	2.0
D kcal/mol	1.2	0.8	0.8	0.8	1.2	0.9	0.9
$\lvert\text{SecStruc}_w\rvert$	5.0×10^6	2.1×10^6	2.5×10^6	1.4×10^6	1.2×10^7	5.5×10^5	1.7×10^6
$\lvert\text{SecStruc}_s\rvert$	1.4×10^7	5.5×10^5	1.2×10^6	4.8×10^5	9.0×10^6	2.2×10^5	1.7×10^6
$\lvert\text{MS}_w\rvert$	2,628	14,943	7,783	6,481	6,291	1,332	238
$\lvert\text{MS}_s\rvert$	19,936	2,308	5,936	3,746	3,850	577	238

Table 4.2 Data returned by `RNAsubopt` and `Barrier`.

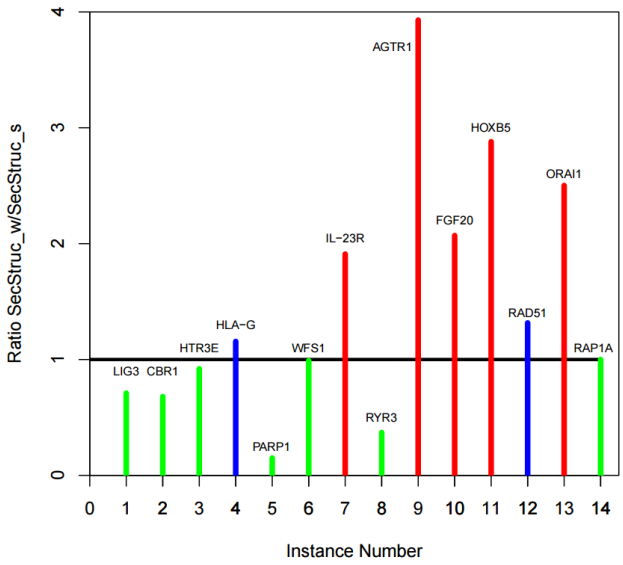

Fig. 4.7 $\lvert\text{SecStruc}_w/s\rvert$ ratios: The green colour indicates a larger or equal number of secondary structures calculated by `RNAsubopt` for the allele with the stronger microRNA binding. For the blue colour the ratio is larger but close to one, and the red colour indicates a much larger number of secondary structures for the allele with the weaker microRNA binding.

computed by `Sfold`, where the calculation is restricted to short base-pair blocks within a 4 nt single-stranded segment of a putative binding site.

The energy cost for the translational and rotational entropy loss is called initiation energy ΔG_{init}, and the standard setting in the STarMir tool is $\Delta G_{\text{init}} = 4.09$ kcal/mol, i.e. $\Delta G_{\text{nucl}} + \Delta G_{\text{initiation}} < 0$ kcal/mol can be seen as a basic requirement for miRNA–mRNA interaction.

(b) $\Delta G_{\text{disrupt}}$ is the energy needed for the disruption of base pairs that are present within a putative binding site in a given mRNA secondary structure. We set $\Delta\Delta G_{\text{dis}} = \Delta G^{w}_{\text{disrupt}} - \Delta G^{s}_{\text{disrupt}}$.

(c) ΔG_{hybrid} is the energy gained by the hybridisation of the miRNA with the particular binding site.

(d) ΔG_{total} is the basic STarMir score defined by $\Delta G_{\text{total}} = \Delta G_{\text{hybrid}} - \Delta G_{\text{disrupt}}$. We set $\Delta\Delta G_{\text{tot}} = \Delta G^{s}_{\text{total}} - \Delta G^{w}_{\text{total}}$.

The binding regions and energy values shown in Table 4.3 are determined by the STarMir prediction with the strongest seed match among all predictions having the SNP position inside, leading in some cases to weaker ΔG_{total} values. In case of a missing strong seed match the target predictions provided by the PITA tool [117] are taken into account.

By $|\text{BP}_w|$ and $|\text{BP}_s|$ we denote the number of base pairs in the corresponding mfe-conformation within the miRNA binding region predicted by STarMir for an individual allele (w = *weak* and s = *strong* allele with respect to miRNA binding). Detailed information about the STarMir output is provided in Appendix B, which also includes PITA predictions and, where available, data provided by FindTar [176]; for target predictions by other tools. Except for IL-23R (L=851), the ΔG_{total} predictions bySTarMir are stronger for the allele identified for stronger miRNA bindings in the corresponding publication (cf. the row for $\Delta\Delta G_{\text{tot}}$ and description of instances in Section 4.6; we recall that for ORAI1 no particular miRNA is mentioned). Except for LIG3 (L=124), HTR3E (L=302), HLA-G (L=386), IL-23R (L=851) and AGTR1 (L=888), the absolute value of $\Delta G_{\text{disrupt}}$ is smaller for the allele with the stronger miRNA binding stated in the corresponding publication (s-allele indicated in Table 4.3, see also the row for $\Delta\Delta G_{\text{dis}}$). For IL-23R and AGTR1, the total STarMir score is positive for both alleles, for IL-23R even with $\Delta G^{s}_{\text{total}} > \Delta G^{w}_{\text{total}}$.

The instance [IL-23R; rs10889677; let-7e] has been analysed in [169], with strong experimental evidence for an inhibitory effect of let-7e on the C-allele. The binding patterns for the C allele provided by STarMir (see appendix B)

and in Figure 3A of [169] differ only slightly towards the 5' end of the C allele. We note that for the A allele we have the only case in Table 4.3 where the SNP position is not within the 'binding site' predicted by `STarMir`. The `PITA` tool returns (the equivalent of) $\Delta G_{\text{total}} = -14.12\,\text{kcal/mol}$ for the C allele, and $\Delta G_{\text{total}} = -10.02\,\text{kcal/mol}$ for the A allele (see appendix B), which is in line with the experimental data from [169]. The absolute value of $\Delta G_{\text{disrupt}}$ is also slightly smaller for the C allele.

For [LIG3; rs4796030; miR-221], the absolute value of $\Delta G_{\text{disrupt}}$ is larger for the C allele by 4.19 kcal/mol, and there is no `PITA` prediction with the SNP position inside the reported binding region. The `PITA` predictions close to the SNP position favour the C allele with respect to the total score and ΔG_{hybrid}, which complies with [164]. The selection of positions 77–98 for the A allele and 80–98 for the C allele is motivated by the seed-like bindings predicted by `STarMir`, see appendix B.

For ORAI1, the positions 69–88 were selected for the C allele due to the larger value of $\Delta G^{s}_{\text{nucl}} = -4.32\,\text{kcal/mol}$.

For [HTR3E; rs56109847; miR-510-5p], the ΔG_{hybrid} values is stronger for the G allele (*s*-case) by 6.2 kcal/mol, and the stronger binding prediction is supported by `PITA` and `FindTar` (no binding prediction for A allele - *w*-case).

For [HLA-G; rs1063320; miR-148a-3p], `PITA` returns a stronger total score for the G allele, which is in line with the `STarMir` total score and experimental data from [167]. The absolute value of $\Delta G_{\text{disrupt}}$ is larger for the G allele (by 1.12 kcal/mol for `PITA` and 1.73 kcal/mol for `STarMir`). `FindTar` also strongly supports the miRNA binding to the G allele.

For [AGTR1; rs5186; miR-155-5p], the ΔG_{hybrid} values is stronger for the A allele (*s*-case) by 4.3 kcal/mol. The stronger binding prediction for the A allele is supported by `PITA` and `FindTar` (no binding prediction for C allele - *w*-case).

For the fourteen instances under consideration, the `PITA` predictions either support (or improve for IL-23R) the `STarMir` predictions displayed in Table 4.3, or no predictions with the SNP position inside the reported binding region are returned. In summary, if `STarMir` and `PITA` (for IL-23R) are taken together, the ΔG_{total} predictions are stronger for the corresponding allele (*s*-case) identified in the description of the fourteen instances in Section 4.6.

	LIG3	CBR1	HTR3E	HLA-G	PARP1	WFS1	IL-23R	RYR3	AGTR1	FGF20	HOXB5	RAD51	ORAI1	RAP1A
L(3' UTR)nt	124	284	302	386	769	779	851	880	888	903	952	978	1034	1078
w-allele	A	G	A	C	C	A	A	G	C	T	G	A	T	C
s-allele	C	A	G	G	T	G	C	A	A	C	A	G	C	A
miR-	221	574	510	148a	145	668	let-7e	367	155	433	7	197	519a	196a
SNP pos	83	133	76	233	607	253	309	839	86	182	141	718	86	366
BindSite-w	77-98	121-62	50-80	221-39	592-614	234-58	291-308	830-57	57-90	166-87	126-54	707-25	81-102	348-70
BindSite-s	80-98	121-62	50-80	221-39	592-614	234-58	291-310	835-57	57-90	166-87	126-54	707-25	69-88	348-70
$\Delta G^{w}_{\text{total}}$	-8.53	-19.72	-14.24	-14.62	1.70	-6.40	0.70	4.35	11.28	-2.18	-8.29	-7.00	5.62	-11.30
$\Delta G^{w}_{\text{disrupt}}$	-11.77	-8.09	-8.26	-9.28	-20.90	-14.20	-21.40	-19.65	-27.89	-10.12	-13.61	-15.00	-23.52	-5.40
$\Delta G^{w}_{\text{hybrid}}$	-20.30	-27.80	-22.50	-23.90	-19.20	-20.60	-20.70	-15.30	-16.60	-12.30	-21.90	-22.00	-17.90	-16.70
$\Delta G^{w}_{\text{nucl}}$	-3.39	-5.36	-2.46	-4.55	-0.65	-1.72	-0.05	-0.19	-2.27	-4.66	-0.49	-1.20	-0.02	-3.97
$\Delta G^{s}_{\text{total}}$	-8.64	-21.84	-16.05	-19.49	-7.51	-15.96	1.12	-3.38	7.66	-5.23	-11.20	-14.31	-7.21	-16.25
$\Delta G^{s}_{\text{disrupt}}$	-15.96	-5.96	-12.65	-11.01	-12.19	-10.64	-25.82	-11.42	-28.56	-9.27	-11.20	-14.29	-9.49	-5.05
$\Delta G^{s}_{\text{hybrid}}$	-24.60	-27.80	-28.70	-30.50	-19.70	-26.60	-24.70	-14.80	-20.90	-14.50	-22.40	-28.60	-16.70	-21.30
$\Delta G^{s}_{\text{nucl}}$	-0.20	-5.22	-5.05	-3.24	-0.93	-4.42	0.00	-1.22	-0.58	-2.36	-1.39	-0.56	-4.32	-6.97
$\Delta\Delta G_{\text{tot}}$	-0.11	-2.12	-1.81	-4.87	-9.21	-9.56	0.42	-7.73	-3.62	-3.05	-2.91	-7.31	-12.83	-4.95
$\Delta\Delta G_{\text{dis}}$	4.19	-2.13	4.39	-12.89	-8.71	-3.56	4.42	-8.23	0.67	-0.85	-2.41	-0.71	-14.03	-0.35
$\lvert BP_w \rvert$	9	14	9	9	22	12	15	14	24	10	15	6	16	12
$\lvert BP_s \rvert$	7	12	15	11	11	12	17	7	22	9	15	6	6	12

Table 4.3 microRNA binding predictions by `STarMir`.

4.7.3 Analysis of meta-stable conformations

MSbind Tool

To analyse subsets of metastable conformations (C) we developed the tool `MSbind` to calculate features of metastable conformations determined by putative microRNA binding sites.

Inputs:

- -f File containing the sequence and metastable conformations in dot-bracket notation sorted by energy.
- -s Target site start position.
- -e Target site end position.
- -c (Optional) Number of metastable conformations to consider from the input file.
- -o Output file name.

The format of the input file requires the sequence be on the first line and remaining lines be formatted as follows: Structure number | Structure in dot-bracket notation | Structure energy | Barrier height. The command `./MSbind -f infile -s 224 -e 232 -o outfile -c 100` thus considers the structural region of nucleotides 224 to 232 of the first 100 metastable conformations in file `infile`. See Appendix A for source code. `MSbind` then:

- Calculates the number of structures with less, equal and greater number of target site pairings in comparison to the MFE structure.
- Sorts and outputs structures by number of target site base pairings in comparison to the MFE target site.
- Calculates the average ΔG energy score and barrier depth over all structures and those with less, equal and greater pairings to the MFE site.

`STarMir` and `PITA` operate on sequences as input, not on representations of secondary structures. Therefore, we utilise `RNAeval` [65] for the energy evaluation of meta-stable conformations within binding regions predicted by `STarMir`, which also complies with the data generated by `RNAsubopt` and `Barriers`, see Table 4.2. In order to facilitate a coherent analysis of energy

values, we employ energy values calculated for binding regions within the corresponding mfe-structure as templates for comparisons, instead of using the associated $\Delta G_{\text{disrupt}}$ values reported in Table 4.3.

Let S denote a secondary structure (either mfe-conformation or meta-stable conformation) for a 3' UTR (wild-type or SNP-type) listed in Table 4.2. By S_{open} we denote the associated secondary structure where all base pair bindings within the miRNA binding region reported in Table 4.3 are removed. For example, if S is the mfe-conformation of the 3' UTR of the C allele of LIG3, then S has seven base pair bindings in positions 80–98 (see Table 4.3), and the seven base-pair bindings are removed in S_{open}. For S being a meta-stable conformation, the number of base-pair bindings within the miRNA binding region can be larger, the same or smaller in comparison to the corresponding value reported in Table 4.3 for the mfe-conformation. We then define

$$\Delta G^{\text{ind}}_{\text{open}} = \texttt{RNAeval}(S) - \texttt{RNAeval}(S_{\text{open}}) \tag{4.1}$$

$$\Delta\Delta G_{\text{ind}} = \Delta G^{\text{ind:}s}_{\text{open}} - \Delta G^{\text{ind:}w}_{\text{open}} \tag{4.2}$$

The index 'ind' specifies the different cases we consider, and the different values in (4.1) and (4.2) relate either to individual structures or to average values (according to the value assigned to 'ind') over sets of meta-stable conformations:

(a) ind = mfe indicates the single mfe-conformation.

(b) ind = tot indicates the average value over all meta-stable conformations as counted in Table 4.2. For example, $\Delta G^{\text{tot:}s}_{\text{open}}$ stands for the average value over 317 meta-stable conformations in case of the C allele of LIG3, see Table 4.2.

(c) ind = N+ indicates for N = 20, 60 and 100 the N+ meta-stable conformations S with the N+ lowest free energy values calculated by $\texttt{RNAeval}(S)$ that are above the mfe conformation. Since each energy level usually adds more than a single conformation, the notion 'N+' indicates that the highest energy level involved covers N conformations above the mfe structure, plus in most cases some more structures, i.e. the actual number N+ of conformations can be slightly larger than N. There are a few exceptions for 60+, where the number is between 50 and 60, be-

cause the next energy level results already in a conformation number above 100. For example, for PARP1, 60+ means 55, because 55 conformations are accumulated at energy level −186.4 kcal/mol, while level −186.3 kcal/mol adds 48 conformations, which leads to 100+ = 103.

Furthermore, we order the meta-stable conformations S_{ms} with respect to the absolute value of $|\Delta G_{\mathrm{open}}|$ and the depth $D(S_{\mathrm{ms}})$ (deepest first), respectively. As in (4.2), we define

$$\Delta\Delta G_{\mathrm{asc:N+}} = \Delta G_{\mathrm{open}}^{\mathrm{asc:N+:}s} - \Delta G_{\mathrm{open}}^{\mathrm{asc:N+:}w}; \tag{4.3}$$

$$R_{\mathrm{N+}} = \mathsf{D}_{\mathrm{N+}}^{w}/\mathsf{D}_{\mathrm{N+}}^{s}, \tag{4.4}$$

where $\mathsf{D}_{\mathrm{N+}}^{w/s}$ denotes the average depth of the N+ deepest meta-stable conformations. The index 'asc' in (4.3) indicates that, unlike in (4.2), the N+ meta-stable conformations are ranked in ascending order with respect to $|\Delta G_{\mathrm{open}}|$, and the average value is taken over N+ conformations.

Finally, we combine 'opening energies', as defined in (4.2) and (4.3), with the depth $D(S_{\mathrm{ms}})$ of meta-stable conformations, as defined by $\mathsf{D}_{\mathrm{N+}}^{w/s}$ and used in (4.4): We look at the average depth of structures S_{ms} with the N+ smallest values of $|\Delta G_{\mathrm{open}}|$. Let $\mathsf{D}_{\mathrm{open}}^{\mathrm{N+:}s}$ and $\mathsf{D}_{\mathrm{open}}^{\mathrm{N+:}w}$ denote the average depth of meta-stable conformations S_{ms} counted in the calculation of $\Delta G_{\mathrm{open}}^{\mathrm{asc:N+:}s}$ and $\Delta G_{\mathrm{open}}^{\mathrm{asc:N+:}w}$.

We then set

$$\Delta\mathsf{D}_{\mathrm{open}}^{\mathrm{N+}} = \mathsf{D}_{\mathrm{open}}^{\mathrm{N+:}s} - \mathsf{D}_{\mathrm{open}}^{\mathrm{N+:}w} \tag{4.5}$$

Comprehensive information about the distribution of meta-stable conformations and their respective energy values is provided in Appendix C. In Table 4.4 we report representative data that are useful for discriminating between the two alleles involved for each instance. Positive values of $\Delta\Delta G_{\mathrm{mfe}}$, $\Delta\Delta G_{\mathrm{tot}}$, $\Delta\Delta G_{\mathrm{N+}}$, $\Delta\Delta G_{\mathrm{asc:N+}}$ and $\Delta\mathsf{D}_{\mathrm{open}}^{\mathrm{N+}}$ are interpreted as being in favour of the allele with the stronger miRNA binding stated in the underlying literature source, which is also the case for $R_{\mathrm{N+}} < 1$.

	LIG3	CBR1	HTR3E	HLA-G	PARP1	WFS1	IL-23R	RYR3	AGTR1	FGF20	HOXB5	RAD51	ORAI1	RAP1A
L(3' UTR)nt	124	284	302	386	769	779	851	880	888	903	952	978	1034	1078
w-allele	A	G	A	C	C	A	A	G	C	T	G	A	T	C
s-allele	C	A	G	G	T	G	C	A	A	C	A	G	C	A
$\Delta\Delta G_{\text{mfe}}$	-2.20	1.00	-13.51	-0.80	13.02	0.00	-4.20	0.75	2.70	3.20	-0.70	-1.90	21.81	0.00
$\Delta\Delta G_{\text{tot}}$	-1.95	1.31	-11.18	-0.80	17.98	0.03	-7.44	0.82	2.74	2.83	-0.56	-1.85	21.80	0.00
$\Delta\Delta G_{100+}$	-2.60	2.19	-11.26	-1.18	12.92	0.00	-4.20	0.58	2.70	3.91	-0.65	-1.90	21.81	0.00
$\Delta\Delta G_{60+}$	-3.20	2.10	-11.39	-1.73	11.48	0.00	-4.20	0.66	2.70	4.19	-0.65	-1.90	21.81	0.00
$\Delta\Delta G_{20+}$	-3.07	3.56	-10.28	-1.61	7.03	0.00	-4.20	0.75	2.70	4.67	-0.70	-1.90	21.81	0.00
$\Delta\Delta G_{\text{asc:}100+}$	-1.90	1.04	-6.86	-0.69	21.84	7.42	-13.40	2.00	2.72	1.84	1.50	-2.18	21.65	0.00
$\Delta\Delta G_{\text{asc:}60+}$	-1.58	0.98	-4.70	-0.78	21.36	7.48	-13.40	2.13	2.72	1.93	3.38	-2.37	21.54	0.00
$\Delta\Delta G_{\text{asc:}20+}$	-0.91	0.93	-4.38	-0.69	21.04	7.62	-13.40	2.61	2.70	2.00	2.36	-2.61	21.00	0.00
R_{100+}	0.96	0.90	1.14	1.00	0.92	1.00	0.98	0.92	1.25	1.06	1.06	1.00	1.14	1.00
R_{60+}	0.93	0.92	1.11	1.00	0.93	1.00	1.00	0.91	1.32	1.05	1.05	1.00	1.10	1.00
R_{20+}	0.88	0.91	1.20	1.01	0.93	1.00	1.00	0.93	1.09	1.04	1.04	1.00	1.07	1.00
$\Delta\text{D}_{\text{open}}^{100+}$	0.05	-0.05	0.15	0.02	0.35	-0.08	0.30	0.04	1.13	-0.07	0.17	0.23	-0.05	0.00
$\Delta\text{D}_{\text{open}}^{60+}$	0.05	-0.11	0.26	0.49	0.53	-0.10	0.40	0.00	1.13	-0.11	-0.14	0.39	0.08	0.00
$\Delta\text{D}_{\text{open}}^{20+}$	0.04	0.00	0.13	-0.02	0.64	-0.16	0.40	-0.26	1.13	-0.15	-0.03	1.05	0.75	0.00

Table 4.4 Energy values calculated by `MSbind` and `RNAeval`.

For LIG3 (L=124), HTR3E (L=302), HLA-G (L=386), IL-23R (L=851) and RAD51 (L=978), the data for $\Delta\Delta G_{\text{mfe}}$, $\Delta\Delta G_{\text{tot}}$, $\Delta\Delta G_{\text{N+}}$ and $\Delta\Delta G_{\text{asc:N+}}$ shown in Table 4.4 are not in favour of the allele with the stronger miRNA binding stated in the underlying literature (HOXB5 with L=952 not included here, please see below).

For LIG3, IL-23R and RAD51 the values of $R_{\text{N+}}$ are either close to or smaller than 1.00, and the values of $\Delta\text{D}_{\text{open}}^{\text{N+}}$ are all positive, suggesting more stable local minima for structures with the smallest absolute value of opening energies for the allele with the stronger miRNA binding.

For HLA-G, the values of $R_{\text{N+}}$ are equal or close to 1.00, and the values of $\Delta\text{D}_{\text{open}}^{\text{60+}}$ and $\Delta\text{D}_{\text{open}}^{\text{100+}}$ are positive. Moreover, the $\Delta\Delta G_{\text{asc:N+}}$-values shown in Table 4.4 are close to zero, and the values of $\Delta\Delta G_{\text{N+}}$ are in the range of $\Delta G^{s}_{\text{disrupt}} - \Delta G^{w}_{\text{disrupt}} = -1.73\,\text{kcal/mol}$ from Table 4.3. Thus, for HLA-G the values of $\Delta G^{s/w}_{\text{hybrid}}$ shown in Table 4.3 seem to be decisive for an assessment of a putative miR-148a-3p $\leftrightarrow$ HLA-G/rs1063320 binding (based upon prediction tools).

For HTR3E, only the values of $\Delta\text{D}_{\text{open}}^{\text{N+}}$ are in favour of the allele with the stronger miRNA binding stated in the underlying literature. Unlike the case of HLA-G, the negative values of $\Delta\Delta G_{\text{N+}}$ and $\Delta\Delta G_{\text{asc:N+}}$ are more substantial, i.e. in terms of absolute values above the range of corresponding values from Table 4.3. Thus, only the values of $\Delta G^{s}_{\text{hybrid}}$, $\Delta G^{s}_{\text{nucl}}$ (both Table 4.3) and $\Delta\text{D}_{\text{open}}^{\text{N+}}$ (Table 4.4) support the binding to the G allele (*s*-case).

For HOXB5 (L=952), the $\Delta\Delta G_{\text{N+}}$ values are negative yet close to zero, and the $\Delta\Delta G_{\text{asc:N+}}$ values support the stronger miRNA binding to the A-allele (*s*-case), which makes the HOXB5 instance different from the five instances discussed above. The $R_{\text{N+}}$ values are close to 1.00, and $\Delta\text{D}_{\text{open}}^{\text{100+}}$ is positive.

For RAP1A (L=1078), the `STarMir` predictions for $\Delta G_{\text{disrupt}}$ and ΔG_{hybrid}, respectively, are very close for both alleles. The SNP at position 366 is located in the middle of a loop (positions 363–369) in both mfe secondary structures, which leads to identical values of $\Delta G_{\text{open}}^{\text{ind:}s/w}$ and related features. Therefore, similar to HTR3E, $\Delta G^{s}_{\text{hybrid}}$ and $\Delta G^{s}_{\text{nucl}}$ from Table 4.3 appear to determine the evaluation of miR-196a$\leftrightarrow$RAP1A/rs6573 bindings.

For CBR1 (L=284), WFS1 (L=779), RYR3 (L=880), AGTR1 (L=888), FGF20 (L=903) and ORAI1 (L=1034), the data for $\Delta\Delta G_{\text{mfe}}$, $\Delta\Delta G_{\text{tot}}$, $\Delta\Delta G_{\text{N+}}$ and $\Delta\Delta G_{\text{asc:N+}}$ (Table 4.4) are all in favour of the allele with the stronger miRNA binding stated in the underlying literature. However, for each of the

five instances at least one of the two parameters $R_{\mathrm{N+}}$ and $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{N+}}$ does not fully support the predicted binding.

For RYR3, only $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{20+}} = -0.26\,\mathrm{kcal/mol}$ is clearly not in favour of the predicted binding. Similarly, the values of $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{N+}}$ do not support the predicted stronger binding for CBR1. For ORAI1, this is the case for the $R_{\mathrm{N+}}$ values and $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{100+}} = -0.05\,\mathrm{kcal/mol}$, but with a relatively strong value of $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{20+}} = +0.75\,\mathrm{kcal/mol}$. For AGTR1, all three values of $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{N+}}$ are clearly in favour of the predicted binding pattern. The instances WFS1 and FGF20 are the only two cases where for all N+ values both parameters do not support the predicted stronger binding. Finally, for PARP1 (L=769) all energy values shown in Table 4.4 support the stronger miRNA binding to the T allele (*s*-case).

We recall that for the instance ORAI1 from [174] no individual miRNA is identified in the literature source. For ORAI1 and miRNA-519a-3p, the two binding sites predicted by `STarMir` intersect only by 8 nt, with a positive value $\Delta G_{\mathrm{total}} = 5.62\,\mathrm{kcal/mol}$ for the T-allele (*w*-case), which suggests that no binding occurs. Although $R_{\mathrm{N+}} > 1.00$ for all cases of N+ considered, we obtain strong positive values for $\Delta\Delta G_{\mathrm{mfe}}$, $\Delta\Delta G_{\mathrm{tot}}$, $\Delta\Delta G_{\mathrm{N+}}$ and $\Delta\Delta G_{\mathrm{asc:N+}}$, along with the `STarMir` predictions ${}_{\Delta}G_{\mathrm{total}} = -7.21\,\mathrm{kcal/mol}$ and $\Delta G_{\mathrm{nucl}} = -4.32\,\mathrm{kcal/mol}$ for the C-allele (*s*-case). Moreover, the data for $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{N+}}$ are: $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{10+}} = 0.89\,\mathrm{kcal/mol}$, $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{20+}} = 0.75\,\mathrm{kcal/mol}$, $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{60+}} = 0.08\,\mathrm{kcal/mol}$, which support a binding of miR-519a-3p to the 3' UTR of ORAI1.

4.8 Conclusions

Out of the fourteen instances we analysed, thirteen instances are sensitive to the parameters $\Delta\Delta G_{\mathrm{ind}}$, $\Delta\Delta G_{\mathrm{asc:N+}}$, $R_{\mathrm{N+}}$ and $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{N+}}$ we introduced in Eqn. 4.1 until Eqn. 5. For RAP1A (L=1078), slightly larger values of δE did not create differences between basic parameters for both alleles and eventually led to an unmanageable size of data for standard desktop computer configurations.

The absence of experimental data about copy numbers of mRNA transcripts considered in the present study prevents the selection of a particular value of $\Delta\mathrm{D}_{\mathrm{open}}^{\mathrm{N+}}$ (or of the other two highlighted parameters), which is why we considered three representative values of N+ simultaneously, without calculating p-values. The upper bound of N=100+ is motivated by the data provided in [146], see Figure 3B there.

	LIG3	CBR1	HTR3E	HLA-G	PARP1	WFS1	IL-23R
L(3' UTR)nt	124	284	302	386	769	779	851
w-allele	A	G	A	C	C	A	A
s-allele	C	A	G	G	T	G	C
`StarMir` (total)	+	+	+	+	+	+	−
N+ selection	20+	20+	20+	60+	20+	20+	100+
$_{\Delta\Delta}G_{\text{asc:N+}}$	+/−	+	−	+/−	+	+	−
$R_{\text{N+}}$	+	+	−	+/−	+	+/−	+
$_{\Delta}\text{D}^{\text{N+}}_{\text{open}}$	+	+/−	+	+	+	−	+

	RYR3	AGTR1	FGF20	HOXB5	RAD51	ORAI1	RAP1A
L(3' UTR)nt	880	888	903	952	978	1034	1078
w-allele	G	C	T	G	A	T	C
s-allele	A	A	C	A	G	C	A
`StarMir` (total)	+	+/−	+	+	+	+	+
N+ selection	100+	20+	20+	100+	20+	20+	20+
$_{\Delta\Delta}G_{\text{asc:N+}}$	+	+	+	+	−	+	+/−
$R_{\text{N+}}$	+	−	−	−	+/−	−	+/−
$_{\Delta}\text{D}^{\text{N+}}_{\text{open}}$	+	+	−	+	+	+	+/−

Table 4.5 Summary of results

The data provided in Table 4.4 indicate that $_{\Delta\Delta}G_{\text{mfe}}$, $\Delta\Delta G_{\text{tot}}$, $\Delta\Delta G_{\text{N+}}$ not necessarily contribute to a deeper insight into miRNA binding patterns to different alleles. In particular, $\Delta\Delta G_{\text{tot}}$ and $\Delta\Delta G_{\text{N+}}$ are related only to free energy values of meta-stable structures, which is why a further discrimination by the depth of meta-stable conformations and the opening energy of binding regions was introduced. Consequently, we focus in the summary of findings presented in Table 4.5 on the values calculated for $\Delta\Delta G_{\text{asc:N+}}$, $R_{\text{N+}}$ and $\Delta\text{D}^{\text{N+}}_{\text{open}}$.

In Table 4.5, the row '`StarMir` (total)' indicates by '+' that the ΔG_{total} score (see Table 4.3) supports the allele with the stronger miRNA binding stated in the underlying literature; '+/−' indicates $0 < \Delta G^{s}_{\text{total}} < \Delta G^{w}_{\text{total}}$. As mentioned above, data about estimations of copy numbers are not available for the mRNA transcripts we consider in the present study. In order to avoid the inclusion of irrelevant data (by averaging or thresholding), we consider a 'best case scenario' for each instance: We select a value of N+ in such a way that the support of the miRNA binding to the allele identified in the underlying literature source (*s*-case) is maximised. In case of multiple N+ values (for eleven instances same pattern as in Table 4.5 for at least two N+), the smallest N+ is selected and named in Table 4.5 in the row 'N+ selection'.

For the selected N+, the Table 4.5 entry is labelled as positive '+', if $\Delta\Delta G_{\text{asc:N+}}$ and $\Delta\text{D}^{\text{N+}}_{\text{open}}$ are positive, respectively, or $R_{\text{N+}} < 1.00$. If the data are inconclusive (equal or close to 0.00 or 1.00), we use '+/−'. For example, for HLA-G we select N+ = 60+ and obtain from Table 4.4 the entries for Table 4.5 as follows: $\Delta\Delta G_{\text{asc:N+}} = +/-$, $R_{\text{N+}} = +/-$, and $\Delta\text{D}^{\text{N+}}_{\text{open}} = +$.

Table 4.5 demonstrates that the combined measure $\Delta\text{D}^{\text{N+}}_{\text{open}}$ defined in (4.5) is the best match to the binding predictions, with two inconclusive and two negative values. The inconclusive value of $\Delta\text{D}^{\text{20+}}_{\text{open}}$ for CBR1 is accompanied by two positive values of the other two parameters, and RAP1A is a special instance due to the SNP location, as discussed in Section 4.7.3. The negative values of $\Delta\text{D}^{\text{20+}}_{\text{open}}$ for WFS1 and FGF20 are accompanied by relatively strong positive values of $\Delta\Delta G_{\text{asc:20+}}$.

We hypothesise that an in-depth analysis of meta-stable conformations based upon parameters such as $\Delta\Delta G_{\text{asc:N+}}$, $R_{\text{N+}}$ and $\Delta\text{D}^{\text{N+}}_{\text{open}}$ can provide useful information for the assessment of putative miRNA-mRNA bindings in the context of single nucleotide polymorphisms. In the literature sources we researched for the current study, the number of genes and microRNAs exposed to experimental analysis is relatively small, yet each analysis is time-consuming and costly. Examining features of meta-stable conformations in a preprocessing

step of wet lab experiments may improve the confidence about expected miRNA-mRNA bindings. We emphasise that for ORAI1 no specific microRNA is identified by [174]. The data we presented support the binding of miR-519a-3p to the 3' UTR of ORAI1 in the region of position 86.

Chapter 5

MicroRNA target prediction based upon metastable RNA secondary structures

The contents of this chapter appear in the following publication:

[10] Ouala Abdelhadi Ep Souki, Luke Day, Andreas A. Albrecht and Kathleen Steinhöfel. "MicroRNA Target Prediction Based Upon Metastable RNA Secondary Structures". *Bioinf. & Biomed. Engineering*, vol. 9044 of LNCS, pages 456 - 467, Springer, 2015 doi: 10.1007/978-3-319-16480-9$_4$5.

5.1 Introduction

MicroRNAs are short non-coding RNAs that post-transcriptionally regulate gene expression by binding to target messenger RNAs. As discussed in detail in Chapter 2 (Sec. 2.4), discoveries over recent years have shown microRNAs play a critical role in many aspects of normal and abnormal biology. And, mapping the target genes of specific microRNAs is currently experimentally challenging because each microRNA typically targets multiple genes. Therefore, computational methods to narrow down potential candidates for experimental validation are of interest. Early microRNA target prediction tools such as `TargetScan` [150] and `miRanda` [105] focus primarily on sequence features such complementarity of a seed region. Other tools such as `RNAcofold` [113] take into consideration thermodynamic stability of the miRNA-mRNA duplex. Advanced methods such as `STarMir` [116] also take

into consideration the secondary structure of mRNA at the suspected binding region.

Some tools avoid intra-molecular base pairing by omitting the computation of folded structures within the monomers and therefore rely almost exclusively on the free energy of the duplex formation. The assumption that the mRNA is in linear form certainly reduces the computational complexity. However, studies suggest that this assumption describes only part of the binding process and that prediction tools can be improved by incorporating the folded structure of the mRNA into the prediction algorithm [117, 177]. In reality, either the binding site must not be involved in any base pairing with other parts of the same mRNA, or there should be an energetic penalty for freeing base pairing interactions within the mRNA in order to make the target site accessible for the binding. This energy cost has to be considered as a part of the total hybridization energy [116]. See Chapter 3 (Sec. 3.4) for further discussion on the challenges of computational microRNA prediction.

In Chapter 4 it was shown analyses of microRNA binding sites over metastable secondary structures, by application of `MSbind`, can in the context of Single Nucleotide Polymorphisms (SNPs) provide useful information for the assessment of putative miRNA-mRNA bindings and differences in allelic expression levels. Namely the parameters $\Delta\Delta G_{\text{asc:N+}}$, $R_{\text{N+}}$ and $\Delta \text{D}^{\text{N+}}_{\text{open}}$. In this chapter, we extend this work and present `RNAStrucTar`, a microRNA prediction tool that analyses putative mRNA binding sites within 3' UTRs over metastable secondary structures. The first stage consists of generating conformations that can be classified as deep local minima from a RNA energy folding landscape. The second stage incorporates duplex structure prediction through sequence alignment and energy computation. Target site accessibility related to different sets of metastable conformations is also taken into account. An overall interaction score computed from multiple binding sites is returned.

The approach is discussed in the context of Single Nucleotide Polymorphisms (SNPs). The reason for testing the prediction tool in the context of SNPs is it provides a way to evaluate prediction scores. Each of the cases considered here have strong experimental expression level results. And, we expect the prediction tool to output better scores for the allele (SNP or wild-type) experimentally reported to have greater microRNA suppression.

5.2 Methods

Our work assumes the existence of multiple active RNA conformations instead of a unique MFE conformation as the single biologically active state. The second basic feature of our approach relates to the presence of multiple copies of each individual mRNA. There are two main stages in `RNAStrucTar`: (1) The first stage is the generation of metastable conformations, and (2) the second stage comprises of miRNA target prediction based upon an energy assessment that incorporates target accessibility related to an input set of secondary structures. A flowchart describing the particular steps is given in Figure 5.1.

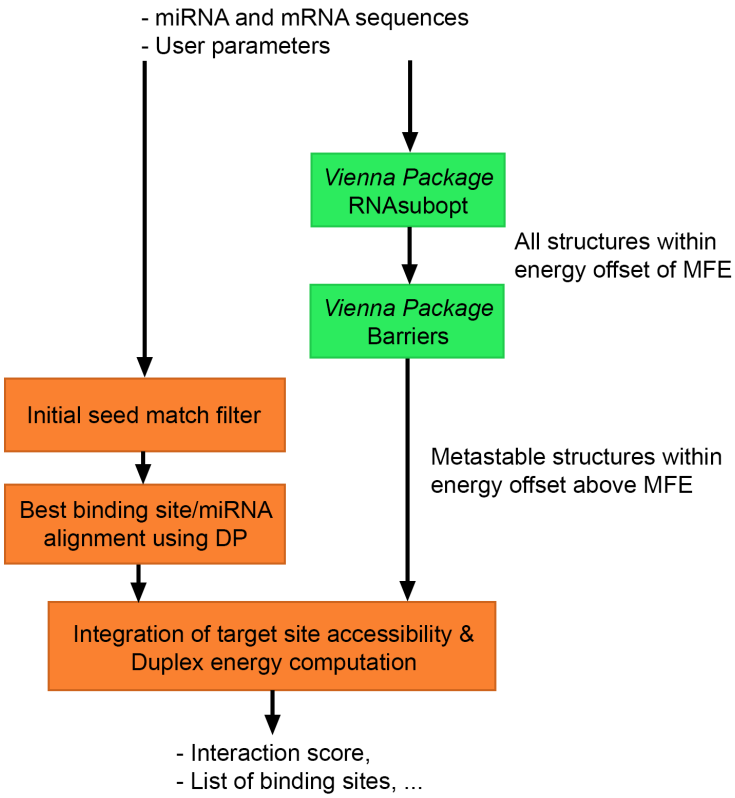

Fig. 5.1 Flowchart of `RNAStructure`.

5.2.1 Metastable secondary structures

Metastable secondary structures are generated by using standard tools provided by the `Vienna RNA package` [65]. The RNAsubopt tool by Wuchty *et al.* [90] generates all suboptimal foldings of a sequence in a partial energy

landscape defined by an energy range ΔG above the minimum-free energy (MFE) structure. A method to elucidate the basin structure of landscapes by means of tree-structures representing local minima and their connecting saddle points is provided by the `Barriers` program [98]. The `RNAsubopt` tool together with the `Barriers` program allows the user to identify the set of metastable conformations MS within an energy range ΔE above the MFE conformation. We denote the number of local minima by $\text{ms} = |\text{MS}|$.

5.2.2 Identification of putative nucleation sites

`RNAStrucTar` first scans the mRNA sequence in search for putative nucleation sites, considering complementariness with the seed region of the miRNA. This seed match step is commonly used and considered as a speed-up factor that accelerates the algorithm while it differs from tool to tool in terms of perfectness of matches. A flexible miRNA seed window - nucleotides 2 to 8, counting from the 5' end of miRNA - is used to scan the mRNA sequence for potential target sites. All results shown in this work were obtained by using 3-mers or 4-mers complementarity to miRNA positions 2-5 in the seed match step. However, the algorithm allows for shorter or longer matches and also matches with varying starting positions.

5.2.3 [binding region, miRNA]-duplex structure prediction

After the seed regions are identified, the upstream flanking region of the seed region is extracted for the next step. Among common features of prediction programs are dynamic programming and the alignment of the miRNA seed region to the target mRNA. We propose a dynamic programming approach for finding minimum energy alignments between the full length of the miRNA and the target sequence for each putative binding site. For this purpose, a modified version of `RNAduplex` [65] is adapted to compute the optimum duplex structure for each putative binding site. For each such site j, the binding pattern and its free energy $\Delta G^j_{\text{binding}}$ are computed according to the seed alignment from the previous step. At the end of this step, weak binding sites are filtered out by applying an energy threshold ϑ with the default setting of -10kcal/mol. This way we obtain k binding sites that satisfy the condition $\Delta G^j_{\text{binding}} \leq \vartheta, j = 1, \ldots, k$.

5.2.4 Integration of target site accessibility

Similar to Kertesz *et al.* [117] and Long *et al.* [116], we adopted the simplifying assumption that the binding of a miRNA to a longer target mRNA should cause a local structural alteration at the target site, but has no long-range effects on the overall target secondary structure. This leads to a breakage of intramolecular bonds within the target region. For each input secondary structure $F_i \in \mathrm{MS}, i = 1,\ldots,ms$ the energy contribution $\Delta G^j_{\mathrm{open},i}$ of the deleted bindings is computed for each j by using `RNAeval` [65] and according to standard of the Nearest Neighbour Model. Given F_i, we denote by $\mathrm{F}_{\mathrm{open},i}$ the associated secondary structure where all base pair bindings within j are removed. We then define

$$\Delta G^j_{\mathrm{open},i} = \texttt{RNAeval}(\mathrm{F}_i) - \texttt{RNAeval}(\mathrm{F}_{\mathrm{open},i}). \tag{5.1}$$

5.2.5 miRNA-target score derived from a single binding site

At this stage, we estimate the free energy of the miRNA:mRNA duplex structure by using `RNAeval`. For each putative binding site j and for each input secondary structure F_i, we generate an artificial RNA sequence which consists of the original mRNA (3' UTR) sequence, a linker sequence XXXX, and the miRNA sequence. The corresponding folding is denoted by $\mathrm{F}^j_{\mathrm{concat},i}$. The score S(miRNA, 3' UTR, F_i, j) $= S_{i,j}$ is then defined by

$$S_{i,j} = \texttt{RNAeval}(\mathrm{F}^j_{\mathrm{concat},i}) - \texttt{RNAeval}(\mathrm{F}_i). \tag{5.2}$$

We integrate the scores of multiple conformations F_i for the $h_j \leq$ ms negative values of $S_{i,j} < 0$ by setting

$$S_j = -\log \sum_{s=1}^{h_j} e^{-S_{i_s,j}}, j = 1\ldots,k. \tag{5.3}$$

See Figure 5.2. In Figure 5.2 we assume for simplicity $h_j =$ ms for all $j \leq k$. The setting according to 5.3 is inspired by the `PITA` scoring function [117].

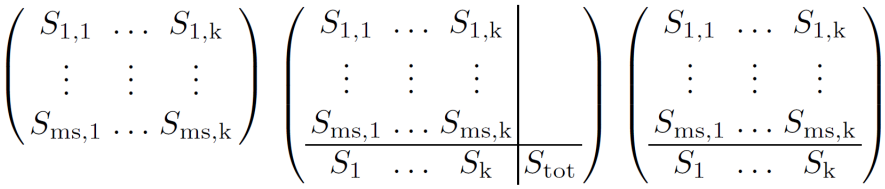

Fig. 5.2 Integration of multiple binding sites and conformations into a single score.

5.2.6 MicroRNA target prediction scores

Some existing target prediction methods check the presence of multiple target sites and take the number of target sites into account for a final score. We explored different ways to account for the occurrence of multiple binding sites and we ended up with a scoring function where the emphasis is on combining strong duplex conformations with a user-defined target region. To integrate multiple sites with S_j-scores for a given miRNA and a fixed 3' UTR into an overall miRNA:target interaction score, we define

$$S_{\text{tot}} = -\log \sum_{j=1}^{k} e^{-S_j}. \tag{5.4}$$

We note that by definition $S_j < 0$ (assuming $h_j \geq 1$) for all $j \leq k$, see Eqn. 5.3. We recall that each $S_j \leq 0$ represents information about $h_j \leq$ ms bindings to metastable conformations $\text{F}_i, i = 1\ldots, h_j$, which justifies the notation S_{tot} as total score. Additionally, other alternative scoring functions were also analysed: For each conformation F_i, the linear sum S_i of $k_i \leq k$ values of $S_{i,j} < 0$ is computed, and for $h \leq$ ms values of $S_i < 0$, the average value is denoted by S_{sum}. Thus, we define

$$S_i = \sum_{t=1}^{k_i} S_{i,j_t}. \tag{5.5}$$

$$S_{\text{sum}} = \frac{\sum_{s=1}^{h} S_{i_s}}{h} \tag{5.6}$$

In addition to S_{tot} and S_{sum}, `RNAStrucTar` allows the user to calculate a score S_u derived from a binding site u that contains a user defined position (usually where the SNP is located) within the input RNA sequence.

$$S_u = \frac{\sum_{s=1}^{h_u} S_{i_s,u}}{h_u}. \tag{5.7}$$

Again, similar to the PITA score [117], we define

$$S_u = \frac{\sum_{s=1}^{h} -\log \sum_{t=1}^{k_i} e^{S_{is},j_t}}{h}. \tag{5.8}$$

We emphasise that based upon Eqns. 5.3 - 5.8 the values of S_j , S_{tot}, S_i, S_{sum}, S_u, and S_P are either negative or not defined (e.g., if $h_j = 0$ for some j or $k_i = 0$ for some i).

5.2.7 Metastable conformations sets

In our analysis presented in Chapter 4, along with the energy ΔE above the MFE conformation, we tried to restrict metastable states to deep local minima. The parameter D indicates the depth of a local minimum or - in other terms - the escape height from a local minimum, which is taken in barrier trees as the distance to the nearest saddle point. We found that out of the three different parameters we introduced, the average depth and the average opening energy of metastable conformations may provide supporting information for a stronger separation between miRNA bindings to the two alleles defined by a given SNP. Here, we aim at individual miRNA-mRNA binding predictions over samples of metastable conformations defined by these parameters. Therefore, we order the metastable conformations with respect to:

(a) The depth $\mathrm{D}(\mathrm{F}_i)$ in descending order (deepest first).

(b) The absolute value of opening energy $\Delta G^u_{\text{open},i}$ of the user defined target region, ranked in ascending order.

We obtain the following two sets, where N is a user defined parameter:

(a) **Set A**: The N deepest metastable conformations among MS.

(b) **Set B**: The N most accessible conformations in the user defined target region among the deepest metastable conformations.

5.3 Results

5.3.1 Test dataset

The tools RNAsubopt and Barriers generate a huge amount of secondary structures, even for a small offset ΔE above the MFE. Consequently, a large scale test or a genome wide prediction analysis is not possible at this stage. In order to test our approach, we use the same data acquisition method as in Chapter 4, i.e. we use miRNA-mRNA pairs from published experimental work where SNPs are linked to specific diseases. SNPs can be located in miRNA binding regions, and consequently they could affect gene expression. `RNAStrucTar` can be used to evaluate how SNPs affect miRNA regulation by using as input the wild type and the SNP variant. Our aim is to determine the ability of `RNAStrucTar` to provide supportive information for a stronger discrimination between miRNA bindings to the two alleles defined by a given SNP (also denoted as RS sequences). The selection of test sequences was governed by the need of having miRNA-mRNA interactions with a high level of experimental validation, for example, by being based upon PCR and/or luciferase reporter assays. We analysed 20 instances of [mRNA/3'UTR; RS; miRNA] interactions, where 14 instances were used in 4 and defined in Table 4.1. The 6 remaining instances were sourced from [178–183]. The sequence IDs were retrieved from the NCBI database and the NCBI Single Nucleotide Polymorphism Database (dbSNP) [126] of nucleotide sequence variation. We also utilised `miRdSNP` [184] and `miRTarbase` [185] for retrieving information related to wild type and variant alleles, ensuring this way a maximum consistency between the publication and the different databases. The sequence length refers to data directly obtained from the NCBI database together with transcript information provided by the ENSEMBL database, and the length ranges between 124 nt and 1167 nt. Some of the publications were sourced from the Human microRNA Disease Database (HMDD) [41]. All results shown in this work were obtained using 3-mers or 4-mers complementarity to the miRNA positions 2-5 in the seed match step. The setting of ΔE depends on the length of the 3' UTR and was selected in such a way that a sufficiently large number of metastable conformations is available. Experimental findings suggest that the typical number of gene copies lies between 5-20, see Section 4.4. Therefore, tests were carried out with N = 10 and N = 20. For each case,

the SNP position was used as the user defined position in order to obtain the score S_u.

5.3.2 Energy scores

We note that the publications of experimental work where the test data are taken from differentiate for each allele pair between weaker bindings (expression levels) and stronger bindings for the miRNA under consideration. Consequently, we calculate energy scores for the weaker and stronger allele, respectively, where it depends on the particular instance which one of the wild type or RS sequence produces the stronger or weaker interaction. Thus, for a given input [mRNA/3'UTR; RS; miRNA], `RNAStrucTar` returns the scores S_{tot} from Eqn. 5.4, S_{sum} from Eqn. 5.6, S_P from Eqn. 5.8, and S_u from Eqn. 5.7 (binding site u contains SNP position) for two alleles, and subsequently the following differences are calculated:

$$\Delta S_{\text{tot}} = S_{\text{tot}}^{\text{stronger}} - S_{\text{tot}}^{\text{weaker}}$$

$$\Delta S_{\text{sum}} = S_{\text{sum}}^{\text{stronger}} - S_{\text{sum}}^{\text{weaker}}$$

$$\Delta S_{\text{P}} = S_{\text{P}}^{\text{stronger}} - S_{\text{P}}^{\text{weaker}}$$

$$\Delta S_{\text{u}} = S_{\text{u}}^{\text{stronger}} - S_{\text{u}}^{\text{weaker}}$$

Negative values of ΔS are expected for a target prediction to be classified as correct. Here, we focus on Case A, although Case B is discussed. The results obtained for the twenty instances and Case A by using 3-mers complementarity regarding the seed match and with setting N = 10 are summarised in Table 5.1. Overall, the score S_{sum} differentiates better than the other scores and it differentiates particularly well between the two alleles on 14 instances, while the S_{sum} scores are indifferent (-1kcal/mol $< \Delta S_{\text{sum}} \leq 0$ kcal/mol.) in four cases (SPI1, IL23R, REV3L, and ORAI1). For the two other cases (HTR3E and FGF20), S_{sum} is in favour of the weaker allele (W-allele). If Case B is taken into account for N = 10, S_{sum} returns a strong and correct prediction for REV3L (Case A is also in favour of the S-allele, but above -1 kcal/mol).

The score S_u gives positive predictions for 12 instances and eight indifferent predictions. However, if S_{sum} and S_u are taken together for Case A, the predictions are in favour of the correct S-allele by at least one of the scores on 16 instances and four indifferent scores.

The two other scores S_{tot} and S_P are in favour of the weaker allele on three instances, and the number of instances where the scores are indifferent is 9 for S_{tot} and 10 for S_P, although a negative value of ΔS is returned on 14 instances for S_{tot} and 13 instances for S_P, but not always below -1 kcal/mol. We conclude that these two scores are less sensitive to binding patterns when compared to S_{sum}.

	LIG3	CBR1	HTR3E	SPI1	HLA-G	MTHFD1	PARP1	WFS1	EFNA1	IL23R
L(3' UTR) nt	124	284	302	369	386	393	769	779	843	851
W-allele	A	G	A	T	C	A	C	A	A	A
S-allele	C	A	G	C	G	G	T	G	G	C
miRNA	221	574	510	569	148a	197	145	668	200c	let-7e
SNP Pos.	83	133	76	330	233	120	607	253	143	309
ΔS_{tot}	+	+	+	0	+	0	0	0	-	+
ΔS_{sum}	+	+	-	0	+	+	+	+	+	0
ΔS_P	+	+	-	0	+	0	0	0	-	0
ΔS_u	+	+	+	0	+	+	+	+	+	0
	RYR3	AGTR1	FGF20	HOXB5	RAD51	REV3L	ORAI1	RAP1A	APP	CD133
L(3' UTR) nt.	880	888	903	952	978	985	1034	1078	1120	1167
W-allele	G	C	T	G	A	C	T	C	C	A
S-allele	A	A	C	A	G	T	C	A	T	C
miRNA	367	155	433	7	197	25	519a	196a	147	135b
SNP Pos.	839	86	182	141	718	460	86	366	171	667
ΔS_{tot}	0	0	-	+	+	0	-	+	+	0
ΔS_{sum}	+	+	-	+	+	0	0	+	+	+
ΔS_P	0	0	-	+	+	0	0	+	+	0
ΔS_u	0	0	0	0	+	+	0	+	0	+

Table 5.1 Summary of predictions by `RNAStrucTar`. '+' ('-') indicates that the score supports the allele reported to have stronger inhibitory effect, with ΔS threshold -1 kcal/mol for '+'. And, '0' means -1 kcal/mol $< \Delta S \leq 0$ kcal/mol.

5.3.3 Comparison to other computational methods

We compare our predictions to those produced by `PITA` [117] and `STarMir` [116]. For the twenty instances we consider, `PITA` returns predictions in favour of the S-allele on 13 instances, with six indifferent scores and one prediction in favour of the W-allele (the seven instances are: PARP1, RYR3, AGTR1, FGF20, RAD51, REV3L, ORAI1). Thus, HTR3E, SPI1, and IL23R are correct by `PITA`, but not by `RNAStrucTar` (S_{sum} only); PARP1, RYR3, AGTR1, and RAD51 are correct by `RNAStrucTar`, but not by `PITA`; both tools fail on FGF20, REV3L, and ORAI1.

The equivalent of S_{sum} for `STarMir` predictions returns score differences in favour of the S-allele on 14 instances, with no indifferent outcomes, but six false predictions on MTHFD1L, EFNA1, IL23R, FGF20, HOXB5, and

RAD51. Thus, HTR3E, SPI1, REV3L, and ORAI1 are correct by STarMir, but not by RNAStrucTar; MTHFD1L, EFNA1, HOXB5, and RAD51 are correct by RNAStrucTar, but not by STarMir; both tools fail on IL23R and FGF20.

In Figure 5.3 we combine the results of the three methods. If we classify an instance as positively predicted if at least two of the methods return a prediction in favour of the S-allele, then 16 correct predictions are made. If only one positive return by a single method is required, then 19 correct predictions are produced by the three methods (only FGF20 is rejected).

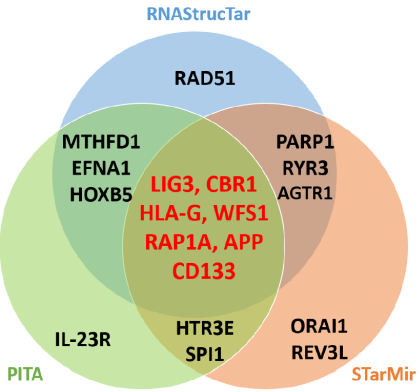

Fig. 5.3 Comparison of predictions between RNAStrucTar, PITA and STarMir.

5.4 Conclusions

We present in this paper RNAStrucTar, a miRNA target prediction tool which incorporates target site accessibility related to metastable secondary structures close to the MFE conformation. We tested our method on 20 miRNA-mRNA interaction pairs that have been experimentally evaluated in the literature. We found that a combination of the two main scores returned by RNAStrucTar supports the experimental findings on 16 instances, with four indifferent outcomes and no false classifications. If STarMir results (14 correct, but partly on different instances) are taken into account, then experimental findings are supported on 18 instances.

Chapter 6

Random vs deterministic descent in RNA energy folding landscape analysis

The contents of this chapter appear in the following publication:

[11] Luke Day, Ouala Abdelhadi Ep Souki, Andreas A. Albrecht and Kathleen Steinhöfel. "Random versus deterministic descent in RNA energy landscape analysis", *Adv. in Bioinf.*, Article ID 9654921, 2016. doi: 10.1155/2016/9654921

6.1 Introduction

Identifying sets of metastable conformations is a major research topic in RNA energy landscape analysis, and recently several methods have been proposed for finding local minima in landscapes spawned by RNA secondary structures. An important and time-critical component of such methods is steepest or gradient descent in attraction basins of local minima. In this chapter, we analyse the speed-up achievable by randomised descent in attraction basins in the context of large sample sets where the size has an order of magnitude in the region of $\sim 10^6$.

6.2 Background

Since Nussinov and Jacobson [59] proposed in 1978 an algorithm to predict a secondary structure with maximum number of base pairs, RNA structure research has expanded into many problems including pseudoknot prediction,

3D structure prediction, the inverse folding problem, modelling of cotranscriptional folding and kinetic folding simulations. The root cause of these problems was stated by Levinthal [81], in the context of protein folding, back in 1969. The conformational space of biomolecules is astronomical yet they manage to fold on a short timescale. In protein folding this paradox has led to the folding funnel hypothesis. Recently, advances in single-molecule experiments have made it possible to observe the transition between folded and unfolded states providing insights into folding times, pathways and energy barriers see [186] and references within. Neupane *et al.* studied the transition paths of nucleic acid and misfolding of a prion protein and found no evidence for a single folding pathway [186].

Flamm and Hofacker provide an overview of the underlying theory and methods for kinetic folding simulations in [187]. Hofacker *et al.* present in [188] the dynamic landscape folding tool `BarMap`. `BarMap` makes use of `RNAsubopt`, `Barriers` and `TreeKin` tools to consider the kinetics of folding in response to environmental changes over macrostates of adjacent landscapes. While basic kinetic moves are addition and deletion of single base pairs, Flamm *et al.* [93] introduced the shift move, which is a combination of a base pair removal and a base pair addition where one position remains invariant. The shift move aims at the simulation of 'defect diffusion' reported in [189], which tries to capture the process where the position of a bulge in a helix may move along a helix as the result of rapid base pair formation and dissociation. Co-transcriptional folding is generally acknowledged as describing the process of how RNA folding happens *in vivo* [190]. RNA is transcribed at a rate of only $\approx$ 30–40 nucleotides per second, where the nascent chain starts folding as soon as it leaves the ribosome. Since helices formed by the incomplete chain may be too stable to refold later on, co-transcriptional folding may drive the folding process to a well-defined folded state that is different from a minimum free energy conformation.

Lorenz and Clote introduce in [191] the $O(n^3)$ time `RNAlocopt` tool for sampling and approximating the total number of metastable conformations using the partition function. And, report the total number of local minima is approximately equal to the square root of the number of structures. However, currently the `RNAlocopt` tool has only been implemented by using the older Turner 1999 energy model without dangling ends.

Li and Zhang [192] focus on the computation of the set of all possible locally optimal stack configurations over the ensemble of putative stacks,

where a new heuristic procedure is utilised for the pathway analysis between local minima. The method targets conformations within a pre-defined energy range above the minimum free energy conformation and aims to identify local minima with high energy barriers. The authors expect the method to be applicable to sequences of up to 250 nt. Huang *et al.* [193] propose a helix-based heuristic for capturing at least significant subsets of local minima of an RNA folding space. Helices are classified by five loop types that are closed by a given helix. The construction of folding pathways utilises dynamic programming that ensures the correct nesting and juxtaposition of structural elements, where a number k of best candidates is considered at each step of the construction of a folding pathway (breadth first search). For fixed values of k, the run-time is estimated by $O\left(k^2 n^3\right)$ energy function evaluations.

Kucharík *et al.* [194] introduce a new connectivity model of attraction basins within energy landscapes, along with the new tool `RNAlocmin` that is designed for generating sets of local minima based upon modified Boltzmann sampling and steepest descent within RNA energy landscapes. To allow for insights into the dynamic behaviour of RNA structure the authors introduce the concept of a Basin Hopping Graph (BHG), where a vertices represent attraction basins with local minima and edges represent energy favourable direct transitions between local minima. As pointed out by the authors, a disadvantage of representing RNA landscapes as a barrier tree is that geometric information is lost. By considering the neighbourhood relation between basins a BHG provides more insight into energy favourable intermediate states. The authors present various comparisons to `RNAlocopt` [191] regarding the coverage of local minima within a given time frame, which turn out to be in favour of `RNAlocmin`, partly with large differences in the number of detected local minima.

While `RNAlocmin` is already relatively fast, we are looking in this work at run-time improvements by randomising the descent within attraction basins. Furthermore, we are interested in the coverage of local minima by deterministic and random descent methods. We note that by using randomised strategies, the completion of steepest descent is not further guaranteed. For large samples even a moderate time improvement of the descent procedure for each individual sample can result in a significant speed-up of the overall processing time. In this work, we take `RNAlocmin` [194] as deterministic steepest descent benchmark method for comparison.

6.3 RNA folding landscapes

Here we briefly reiterate, the energy landscape of an RNA sequence R, denoted by $\mathsf{L}(R) = [\mathsf{C}, \mathsf{N}, E]$, can be described by three components: a set of secondary structure conformations C, a neighbourhood function N and a free energy evaluation function E (see defn. 3.1). The conformation space C consists of secondary structures, and computed by tools such as `RNAfold` or `Mfold`. It is also important to distinguish between two types of conformation spaces: non-canonical and a more restricted canonical spaces where isolated base pairs are admitted (see defn. 2.4). Here, we consider canonical conformation spaces only.

The neighbourhood function N_S of a secondary structure S defines the adjacency of the conformation space C. For the secondary structure S, its neighbourhood N_S is a set of structures that are reachable from S by applying a single operation from a move set, $S \rightarrow S' \in \mathsf{N}_S$. Flamm *et al.* [93] describe two move sets for RNA folding, a basic move set consisting of insertion and deletion of base pairs, and a move set where a shift move to facilitate chain sliding is included. In the present work, we consider the insertion and deletion move set, with the reason being that the `Barriers` implementation of the two move sets generates shift moves only for non-canonical structures. The basic move set is therefore defined in the following way:

(1) Single or double insertion:

   (a) A single base pair may be inserted at position (i, j), if an existing helix is extended; that is, $(i+1, j-1)$ and/or $(i-1, j+1)$ are paired.

   (b) Two base pairings may be inserted at positions (i, j) and $(i+1, j-1)$, if i or j is not adjacent to an existing base pair belonging to the same helix; that is, $i-1$ and $i+2$ or $j+1$ and $j-2$ are unpaired.

(2) Single or double deletion:

   (a) A single base pairing (i, j) may be deleted, if its removal does not result in a non-canonical structure.

   (b) Two base pairings (i, j) and $(i+1, j-1)$ may be deleted,

      i. if position $i-1$ and $i+2$ are unpaired,

ii. if position $i-1$ is the closing base of a different helix and $i+2$ is unpaired.

iii. if $j+1$ is unpaired and $j-2$ is unpaired,

iv. if $j+1$ is the opening base of a different helix and $j-2$ is unpaired.

Additionally, the moves must also satisfy the secondary structure rules, namely, minimum hairpins of length 3 and no pseudoknots. The number of possible neighbours is bounded by $O(n^2)$, where n is the length of the structure. The implementation `RNAbor` for studying statistics of RNA structural neighbours has been introduce in [195]. `RNAbor` computes the number and Boltzmann probabilities of all structures having base pair distance d to a input structure S. `RNAbor` uses dynamic programming and has a complexity of $O(n^4)$. Currently, `RNAbor` works for non-canonical neighbour spaces and uses an older version of the Nearest Neighbour energy model.

The energy function $E : \mathsf{C} \to \mathbb{R}$ calculates the free energy of secondary structures and can be calculated by using, for example, the `RNAeval` tool. Finally, a structure $S_m \in \mathsf{C}$ is metastable (or a local minimum) of the landscape if all its neighbours have higher or equal energy, i.e. $\forall S(S \in \mathsf{N}_S \to E(S) \geq E(S_m))$.

6.3.1 Main features of `RNAlocmin`

Here, we briefly describe the main features of `RNAlocmin` as presented in [194]. The `RNAlocmin` tool accepts as input a set $\{S\}$ of RNA secondary structure conformations, and calculates for each structure S its corresponding local minimum conformation that defines the attraction basin to which S belongs. The underlying method implemented by `RNAlocmin` is a descent algorithm. `RNAlocmin` implements three types of descent: (1) a gradient or steepest descent, (2) a first-lower descent, and (3) a random first-lower descent. Along with the local minima structures S_m and their free energies $E(S_m)$, `RNAlocmin` counts the total number $c(\{S\}, S_m)$ of input structures S that fold into each particular local minimum S_m. As the number of input structures is typically much larger than the number of local minima, some local minima must be reached by multiple input structures. The values $c(\{S\}, S_m)$ therefore provide some insight into the number of structures belonging to attraction basins of the energy landscapes, and consequently the potential size of those basins.

The input conformations are converted into a numerical representation, where for each base pairing (i, j) the opening position i is stored at its closing position j, and the closing position is stored at its opening position. All unpaired positions are set to 0. For example, the structure `.(((...)))..` of length 12 is represented numerically by

i	1	2	3	4	5	6	7	8	9	10	11	12
Structure	.	(	(	(	.	.	.	)	)	)	.	.
$S[i]$	0	10	9	8	0	0	0	4	3	2	0	0

The numerical representation supports the efficient search for potential closing positions j of an unpaired open position i. Figure 6.1 illustrates typical scenarios for finding a suitable j-position, given position i: (A) The search starts from $j = i+1$. If $S[j] > j$, then j is the first position of a helix and j is updated to $S[j]+1$. For example, as for the structure `.(((...)))..`, if $i = 1$ and $j = 2$, then $S[2] = 10$ and j updated to 11; (B) Position j is the closing of a base pairing within a hairpin region where $S[j] < j$.

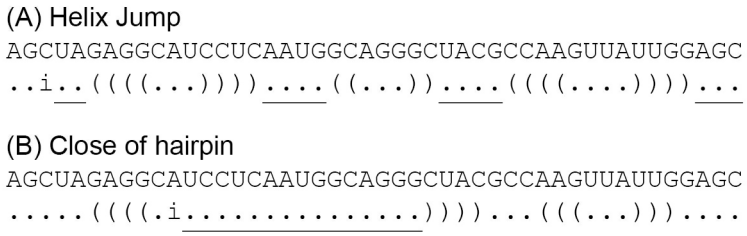

Fig. 6.1 Search for valid base pair (i, j) positions. (a) By using the numerical representation of secondary structure, it is possible to jump over helices in the search for valid j positions. (b) If searching within a hairpin of a helix, then the search can be terminated once a closing bracket is found.

As indicated in the figure, insertion checks only positions where a potential pairing is possible according to the current structure. In the first case (A), a base pair cannot be inserted between $i+1$ and $S[j]$ for a number of values j, i.e., the search for a suitable j 'jumps over helices'. The second case (B) occurs if i is within the hairpin region of a helix, which is recognised from $S[j] < j$.

Like for the `Barriers` tool, it is possible to generate canonical local minima by using `RNAlocmin` through enabling an optional no-loose-pairs parameter (-noLP). Also like `Barriers`, if the -noLP parameter is enabled,

then shift moves are not generated. It is important to note that the canonical neighbourhood generated by `RNAlocmin` differs slightly from that generated by `Barriers`. The neighbourhood generated by `RNAlocmin` is larger than the `Barriers` neighbourhood, because it admits double insertion or deletion of base pairings, if both i and j are adjacent to a pairing. More specifically, the `RNAlocmin` implementation of the double insertion move considers both potential inner and outer pairings. For example, if positions (i, j), $(i-1, j+1)$ and $(i+1, j-1)$ of the structure `..((.i..((...)).j.))..` can form valid pairings, then `RNAlocmin` evaluates both possibilities:

```
(1) ..(((i..((...)).j)))..
(2) ..((.i(.((...)))j.))..
```

However, the outer double insertion move `(1)` is not valid according to the basic move set rules defined above, and it is not generated by `Barriers`. Considering both inner and outer double insertion results in some neighbours being evaluated twice. Additionally, `RNAlocmin` admits double deletion where both i and j are adjacent to a pairing that will not be removed, e.g. `..((i(...)j))..` $\rightarrow$ `..(((.....)))..` is generated by `RNAlocmin`, but not by `Barriers`.

For a sample M of input secondary structures, the time complexity to calculate local minima by using `RNAlocmin` is $O(M \times kn^2 E_n)$, where E_n is the complexity of energy evaluation and k is the maximum number of descent steps to a local minimum. `RNAlocmin` offers two choices for energy evaluation: `energy_of_structure()` and `energy_of_move()`. The energy of the structure method `energy_of_structure()` is equivalent to calling the `RNAeval` tool with time complexity $E_n = O(n)$. The energy of move method, `energy_of_move()` is a local energy update procedure that was introduced in version 2.1.0 of the *Vienna RNA Package* and has time complexity $E_n = O(1)$, achieved by three lookup procedures from the tabulated energy model [194].

6.4 Descent procedures

Here, we describe three descent algorithms implemented by `RNAlocmin` and their modification that make them compatible to the canonical local minima produced by the `Barriers` tool. In particular, the insertion and deletion move functions implemented by `RNAlocmin` were changed according to the move set described in Section 6.3.

Gradient descent

The gradient or steepest descent algorithm calculates and evaluates on each iteration the free-energy of *all* neighbouring conformations reachable from some structure S by insertion or deletion of base pairs. The input conformation is firstly evaluated using the `energy_of_structure()` function, and then a search for neighbouring moves is performed.

If a position i is unpaired, then a search is conducted for valid closing positions, such that (i, j) satisfies the move set conditions described previously in Section 6.3 and Section 6.3.1. When a valid pairing position is found, then its energy is evaluated by using `energy_of_move()`. If the energy returned is lower than all previously seen structures, then the structure is remembered. If a position i is paired, then the pairing $(i, S[i])$ is deleted in case it does not violate the move set conditions. Each iteration continues from the lowest found free-energy structure, or steepest neighbour, until a local minimum is found.

First-lower descent

First-lower descent simplifies the gradient descent by searching for the first energy improvement: The neighbours of the current secondary structure are evaluated by starting from position $i = 1$ of the current secondary structure until a lower energy neighbour is found. Consequently, whenever a lower energy neighbour is found, the search restarts from position $i = 1$ of the lower energy neighbour until a local minimum is found.

Random first-lower descent

In `RNAlocmin`, random first-lower descent works by, on each iteration, generating and storing all neighbour transition moves according to the `RNAlocmin` description in Section 6.3.1, i.e. all potential (i, j) pairing or deletion positions are stored. The list of moves is then randomly shuffled and the shuffled list of moves is evaluated until a lower energy move is found. If no move from the list results in lower energy, then a local minimum has been found. However, this random first-lower descent is implemented only for non-canonical structures within the `RNAlocmin` framework. We implemented a modified random first-lower descent procedure for dealing with canonical structures. The new random descent works by starting the search from a random position, i, of the current structure. Whenever a lower energy move is found between i

and $j \geq i+1$, the structure is updated with the move and the search restarts from another random position i in the updated structure. If no lower energy neighbour is found, then the search restarts from position $i = 1$, which means the current structure is tested for being a local minimum.

Pseudocode 1: Canonical insertion

```
S ⟵ numerical representation of a conformation
LM ⟵ S
min, e ⟵ energy_of_structure(S)
for i = 1 to sequence length do
    if S[i] = 0 then
        for j = i + 1 to sequence length do
            if S[j] > j then
                j ⟵ S[j] (Jump helix)
            if S[j] < j then
                break (Close of hairpin)
            if j − i > 3 and (S[i], S[j]) legal pairing then
                if S[i − 1] or S[i + 1] and S[j − 1] or S[j + 1] ! = 0 then
                    e ⟵ e + energy_of_move(S, i, j)
                    if e < min then
                        S′ ⟵ S
                        S′[i] ⟵ j
                        S′[j] ⟵ i
                        LM ⟵ S′
                        min ⟵ e
                else if (j − 1) − (i + 1) > 3 and S[i + 1] = 0
                and S[j − 1] = 0 and (i, j) a legal pairing then
                    if S[i − 1] < i and S[i + 2] = 0 or S[i − 1] = 0 and
                    S[i + 2] = 0 or S[j + 1] = 0 and S[j − 2] = 0 or
                    S[j + 1] > j and S[j − 2] = 0 then
                        e ⟵ e + energy_of_move(S, i, j)
                        e ⟵ e + energy_of_move(S, i + 1, j − 1)
                        if e < min then
                            S′ ⟵ S
                            S′[i] ⟵ j
                            S′[j] ⟵ i
                            S′[i + 1] ⟵ j − 1
                            S′[j − 1] ⟵ i + 1
                            LM ⟵ S′
                            min ⟵ e
```

Canonical insertion.

Pseudocode 2: Canonical deletion

```
for i = 1 to len do
    if S[i] > S[S[i]] then
        if deletion of i and j does not create a lonely pair then
            e ⟵ e + energy_of_move(S, i, j)
            if e < min then
                S′ ⟵ S
                S′[i] ⟵ 0
                S′[j] ⟵ 0
                LM ⟵ S′
                min ⟵ e
        else if deletion of i + 1, j − 1 does not create a lonely pair then
            e ⟵ e + energy_of_move(S, i, j)
            e ⟵ e + energy_of_move(S, i + 1, j − 1)
            if e < min then
                S′ ⟵ S
                S′[i] ⟵ 0
                S′[j] ⟵ 0
                S′[i + 1] ⟵ 0
                S′[j − 1] ⟵ 0
                LM ⟵ S′
                min ⟵ e
```

Canonical deletion.

6.4.1 RNA sequences

Ten 3' untranslated region (UTR) sequences where identified such that their lengths allow for adequate generation of partial energy folding landscapes. Table 6.1 provides information on the ten human 3'UTR sequences identifed from the NCBI and Ensembl databases.

The partial energy landscape of each sequence was generated by using the *Vienna RNA Package* tools `RNAsubopt` (version 2.1.7) and `Barriers` (version 1.5.2). Table 6.2 shows the total number of structures, $|\mathsf{C}_{\delta E}|$, and local minima, ν, generated by using `RNAsubopt` and `Barriers` within an energy offset δE of the MFE conformation.

The energy offsets of partial landscapes where chosen in such a way that the total number of conformations generated by `RNAsubopt` is between 11×10^6 and 16×10^6. For example, a comparable number of $\sim 15 \times 10^6$ conformations is the output generated by `RNAsubopt` for five instances: GMEB1, LIG3, HTR3E, HLA-G and ALDH4A1. However, the ratio of local minima in the conformation space $|\mathsf{C}_{\delta E}|/\nu$ is, for example, for HLA-G = 17.4 and for

No	Gene Name	ℓ	NCBI Ref. No.	Transcript ID
1	PAX7	99	NM_002584.2	ENST00000375375
2	OXT	99	NM_000915.3	ENST00000217386
3	GMEB1	113	NM_024482.2	ENST00000361872
4	LIG3	124	NM_002311.4	ENST00000262327
5	CBR1	284	NM_001757.2	ENST00000290349
6	HTR3E	302	NM_001256614.1	ENST00000360323
7	HLA-G	386	NM_002127.5	ENST00000360323
8	ALDH4A1	400	NM_170726.2	ENST00000290597
9	MRPL9	407	NM_031420.2	ENST00000368830
10	AQP5	504	NM_001651.3	ENST00000293599

Table 6.1 3' UTR Sequences, ℓ denotes the length of sequences (number of nucleotides).

GMEB1 = 34.0, i.e., GMEB1 has over twice the number of local minima for a comparable total number of conformations.

No	Gene Name	ℓ	δE	$\lvert\mathsf{C}_{\delta E}\rvert$	v	$\lvert\mathsf{C}_{\delta E}\rvert/v$
1	PAX7	99	16.2	14,340,878	50,861	282.0
2	OXT	99	15.0	14,164,430	74,426	190.3
3	GMEB1	113	10.5	15,845,050	466,093	34.0
4	LIG3	124	13.0	15,525,022	317,284	48.9
5	CBR1	284	6.0	10,987,435	643,999	17.1
6	HTR3E	302	9.0	15,095,701	533,316	28.3
7	HLA-G	386	4.2	15,791,146	906,393	17.4
8	ALDH4A1	400	5.4	15,186,200	540,609	28.1
9	MRPL9	407	6.2	14,023,048	41,979	334.0
10	AQP5	504	5.5	11,173,352	714,812	15.6

Table 6.2 Partial energy landscapes; ℓ denotes the length of sequences (number of nucleotides), δE is the energy offset above the MFE structure, $|\mathsf{C}|$ is the number of secondary structures within the partial energy landscape defined by δE, and v is the number of local minima within $|\mathsf{C}|$ identified by `RNAsubopt` and `Barriers`.

6.5 Results

Firstly, we compare the performance of the three descent algorithms in terms of run-time performance and number of observed local minima. Then, we

examine how the different descent methods affect the quality of approximations of the number of local minima.

6.5.1 Run-time and observed local minima

In order to evaluate the three descent procedures, M initial conformations were randomly selected from the top quarter of the energy-sorted partial landscapes, with a subsequent calculation of local minima by using the modified `RNAlocmin` (version 1.0) tool. Structures randomly selected from the highest energy region allow us to sample conformations belonging to a multitude of basins within the partial energy landscape. Table 6.3 shows the percentage of local minima found by each descent method for a random set of conformations in comparison to the number of local minima returned by `Barriers`.

No	Gene	ℓ	ν_{obs}	$M \times 10^6$	Gradient lm(M)	Random lm(M)	First lm(M)
1	PAX7	99	50,861	1.0	63.55	72.64	74.53
2	OXT	99	74,426	1.0	61.74	67.41	72.70
3	GMEB1	113	466,093	2.0	54.54	56.54	57.76
4	LIG3	124	317,284	1.0	42.72	45.50	49.12
5	CBR1	284	643,999	1.5	49.40	48.95	52.07
6	HTR3E	302	533,316	1.5	42.42	44.34	43.55
7	HLA-G	386	906,393	2.5	51.49	52.92	52.99
8	ALDH4A1	400	540,609	1.5	51.06	46.94	51.69
9	MRPL9	407	41,979	1.0	60.44	61.08	61.19
10	AQP5	504	714,812	2.0	58.02	57.98	58.72
				Average	53.54	55.43	57.43

Table 6.3 Observed local minima: Percentage of local minima found by each descent procedure for the same set of M conformations (i.e. last three columns in %).

Over all ten cases, gradient descent results in the smallest number of observed local minima, except for three cases: CBR1, ALDH4A1 and AQP5. For these three cases random descent folds into a slightly smaller number of local minima with a maximum difference compared to gradient of 4.12% for ALDH4A1. This difference suggests that at least for a small number of conformations random descent can take a different folding pathway to a different, possibly lower energy, local minimum compared to gradient descent.

No	ℓ	M $\times 10^6$	`RNAsubopt` time	`Barriers` time	Gradient time	Random time	First time
1	99	1.0	1.48	6.43	5.42	4.03	3.40
2	99	1.0	1.40	5.78	3.03	3.67	3.62
3	113	2.0	1.48	13.63	3.95	6.68	3.38
4	124	1.0	1.70	10.52	3.62	3.27	3.12
5	284	1.5	5.87	18.52	7.22	6.45	6.18
6	302	1.5	6.58	33.90	7.12	6.04	5.93
7	386	2.5	9.25	60.12	12.67	11.10	10.57
8	400	1.5	10.25	53.30	6.62	5.82	5.27
9	407	1.0	15.12	40.12	5.18	4.15	3.95
10	504	2.0	15.75	47.63	9.10	8.40	7.92
		Total	68.88	289.84	63.94	59.61	53.34

Table 6.4 Run-time in minutes of `RNAsubopt`, `Barriers` and modified `RNAlocmin` descent procedures for M conformations. Note, `RNAlocmin` energy evaluation using `energy_of_move()`.

First-lower descent displays the largest number of local minima. The overall average difference is 3.89% for first-lower and 1.89% for random first-lower compared to gradient descent.

Figure 6.2 shows the percentage of run-time improvement of random first-lower and first-lower compared to gradient descent, see also Table 6.4 for the corresponding absolute values [1]. Also included in Figure 6.2, and in Table 6.5, is the average number of gradient descent iterations. For sequences, such as AQP5 with $M = 2 \times 10^6$ samples, where the improvement in run-time is relatively small, the average number of descent iterations is also small. For shorter sequences, such as PAX7 with $M = 10^6$ samples, the average number of iterations is larger. Thus, for our dataset, the run-time improvement suggests a correlation to the number of gradient descent iterations.

Since the energy offsets were chosen in such a way that instances have a comparable number of conformations within their respective partial energy landscapes, a larger subset of the complete energy landscape is considered for shorter sequences. An underlying principle of energy-driven RNA folding is that base pairings stabilise conformations. A secondary structure is said to be saturated, if it is not possible to insert a base pairing without violating the rules of secondary structures. As a larger portion of the full energy landscape

[1] All runtimes generated using Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz and 32GB RAM.

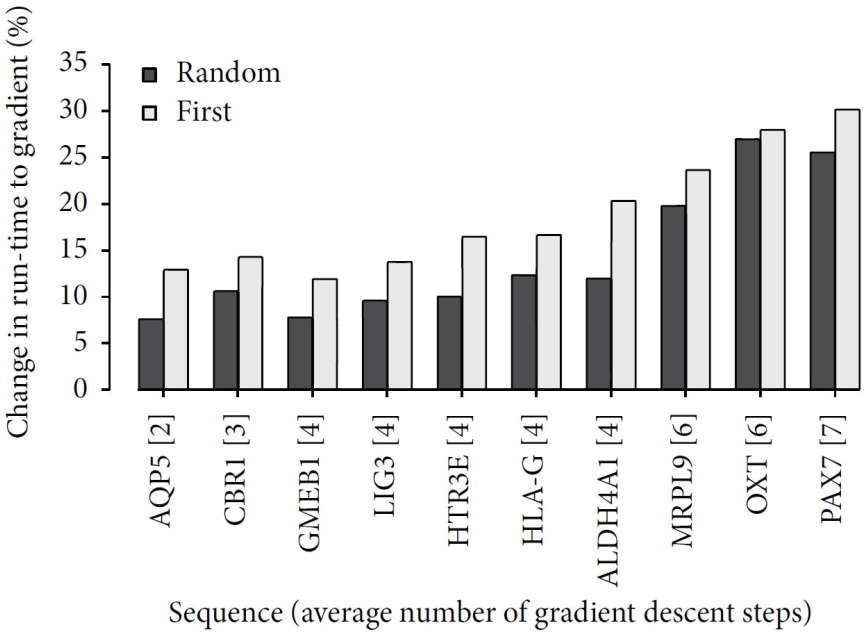

Fig. 6.2 Speed-up in % relative to gradient descent in run-time of first and random first-lower descent for $M = 500$.

Gene	Gradient	Random	% change
AQP5	2	4	7.69
CBR1	3	5	10.66
GMEB1	4	6	7.86
LIG3	4	6	9.67
HLA-G	4	5	12.39
HTR3E	4	5	10.11
ALDH4A1	4	6	12.08
MRPL9	6	7	19.88
OXT	6	8	27.04
PAX7	7	9	25.65

Table 6.5 Descent iterations: Average number of iterations for $M = 500$ and speed-up (in %) in random and first-lower compared to gradient descent.

is generated for shorter sequences, the top quarter of the energy-sorted partial landscape will consist of a larger number of unsaturated conformations. The comparison of descent methods for saturated structures is unlikely to lead to any considerable differences in the run-time, because the cost of deleting base pairings is equal for each descent method. However, for unsaturated

structures the more time-expensive insertion operations are required for the folding process into local minima.

δE	$\lvert C_{\delta E}\rvert$ $\times 10^6$	Gradient time	Random time	First time	Gradient lm(M)	Random lm(M)	First lm(M)
14.0	7.3	14.80	10.65	11.60	45,656	46,005	48,138
16.0	26.7	15.47	10.98	11.48	65,923	70,003	74,939
18.0	86.4	15.83	11.61	14.55	87,307	96,911	106,807
20.0	248.7	17.27	11.47	14.85	108,334	124,566	141,783
22.0	638.5	19.08	11.82	11.78	125,422	150,296	176,904
24.0	1,466.5	21.61	12.17	12.80	138,447	173,275	210,951
26.0	3,018.7	22.28	12.48	12.20	147,261	191,784	241,460

Table 6.6 Increasing energy offset: number of observed local minima and run-time time (minutes) for increasing energy offset for gene OXT. $M = 3 \times 10^6$ randomly sampled from the full partial landscape.

The run-time correlation to descent iterations suggests that random first-lower and first-lower descent are likely to perform particularly well for unsaturated structures. Figure 6.3 shows the run-time difference in percentages for random first-lower descent compared to gradient descent for increasing values of energy offsets. We note that for this analysis the M samples were randomly selected from the unsorted partial conformation space as returned by `RNAsubopt`; see also Table 6.3 for absolute values.

The reason for the selection of M samples from the entire partial space, instead from the highest energy region, is due to the large number of conformations. For example, the number of conformations for offset 26.0 in Figure 6.3 is just over 3 billion; see Table 6.2. The sorting procedure implemented in `RNAsubopt` is memory-expensive, and therefore offsets resulting in very large numbers of conformations exceed the standard desktop computer memory range. In general, a significant run-time improvement is likely to be achieved when folding process proceeds from higher energy conformations within the partial energy landscape.

As can be seen from Table 6.3, first-lower descent and random first-lower descent detect on average for the datasets considered more local minima compared to gradient descent (57.43% and 55.43% compared to 53.54%). Moreover, for all ten partial energy landscapes and the selected values of M, either first-lower descent or random first-lower descent detects more local

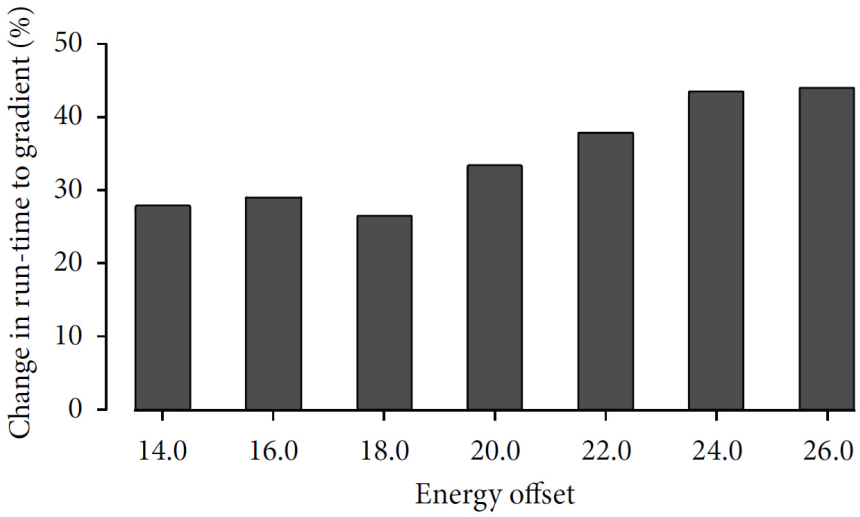

Fig. 6.3 Increasing energy offset: Percentage change in run-time of random descent compared to gradient for OXT and $M = 3 \times 10^6$ (speed-up in % relative to gradient descent).

minima than gradient descent. On the other hand, the run-time is shorter on average and, except for OXT with $M = 10^6$, on all sequences, see Table 6.4.

Figure 6.4 displays the coverage of local minima by different descent methods for PAX7 with $M = 10^6$. The energy values of local minima are rounded to integer values. As can be seen from the upper part, the coverage complies with the `Barriers` data for low energy values up until -4 kcal/mol. The differences in higher values are clearly the result of the random selection of M sample structures. Figure 6.5 provides information about the distribution of sample structures within attraction basins: The left hand side indicates the number of samples (out of M) 'attracted' by local minima of a certain energy value. The figure shows that gradient descent is steering many samples into low energy local minima, whereas first-lower descent and random first-lower descent cover a wider range of local minima.

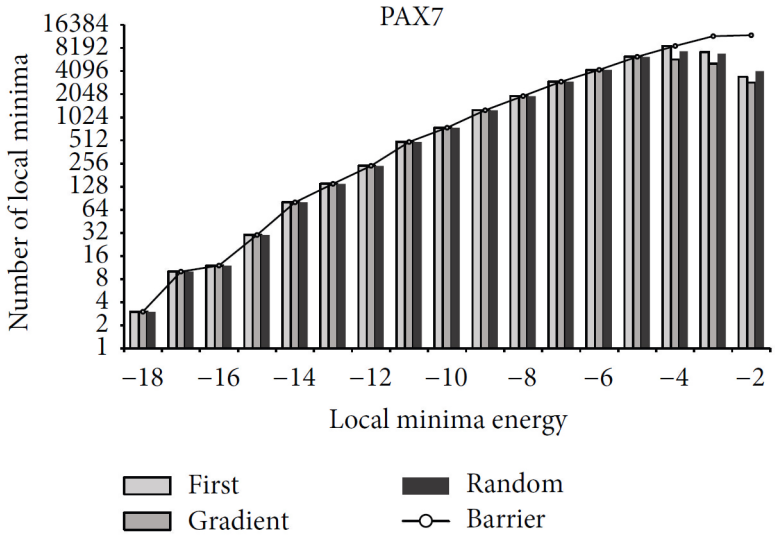

Fig. 6.4 PAX7 local minima coverage.

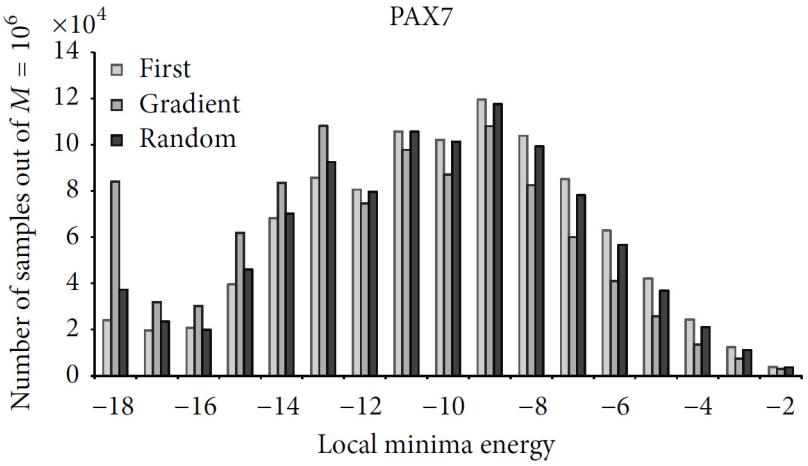

Fig. 6.5 Distribution of local minima by energy.

6.6 Conclusions

In this work, we applied three descent methods to partial RNA energy landscapes and compared run-time and coverage of local minima on random sample sets of conformations taken from the partial energy landscapes induced by ten RNA sequences. While the gain for each individual sample might be marginal, the overall run-time improvement can be significant. In comparison to gradient descent, we obtained on average a total run-time improvement of about 16.6% along with an increase of 7.3% in observed local minima for first-lower

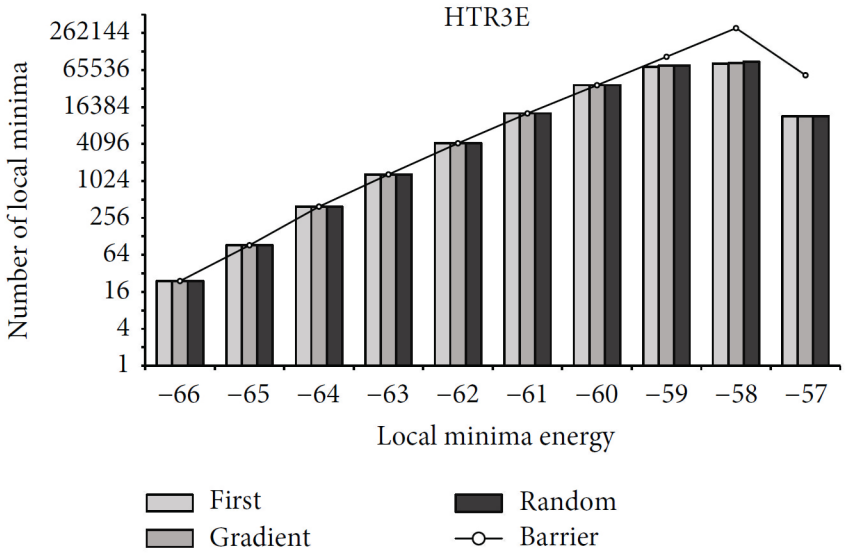

Fig. 6.6 HTR3E local minima coverage.

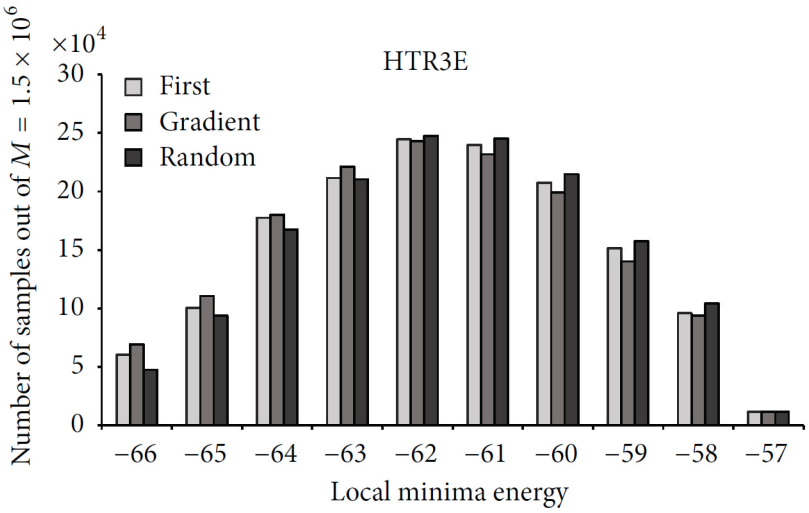

Fig. 6.7 Distribution of local minima by energy.

descent, and a shorter run-time of 6.8% on average with 3.5% more observed local minima for random first-lower descent. One of our main observations is that for all three descent procedures the coverage of local minima produced by `Barriers` is very high for energy values close to the minimum free-energy structure and up until the the region where the samples are randomly selected within the partial energy landscapes. For the large sample size we selected for descent procedures, the coverage of local minima is very high up to energy

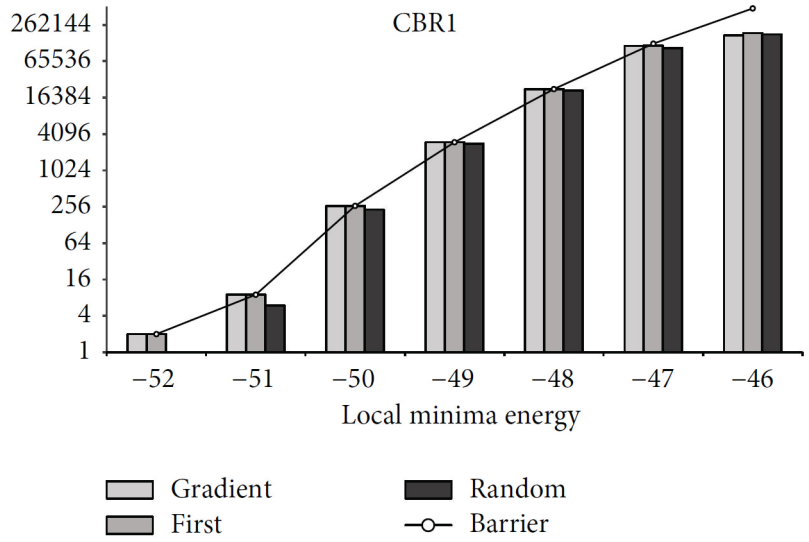

Fig. 6.8 CBR1 local minima coverage.

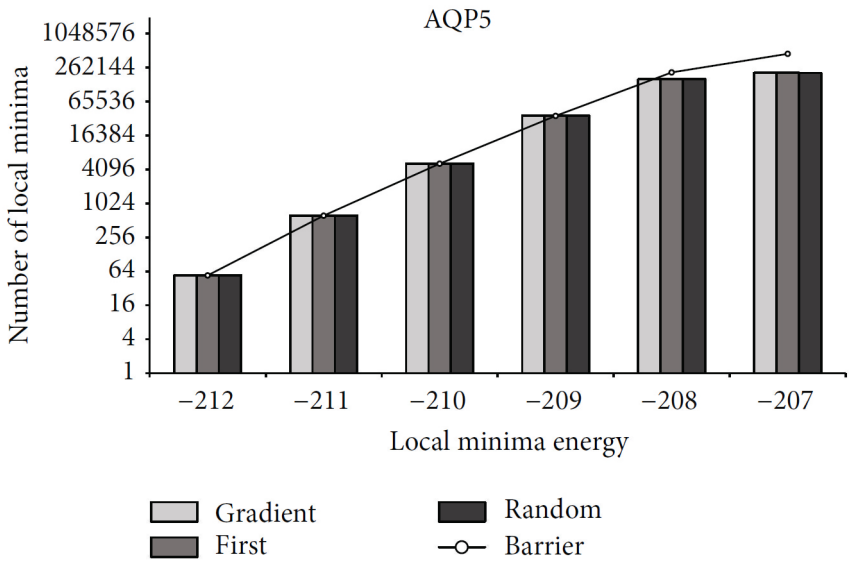

Fig. 6.9 AQP5 local minima coverage.

values of the region where the samples were randomly selected from the partial energy landscapes, i.e., the difference to the total set of local minima is mainly due to the upper area of the energy landscapes. It is an open question if, for highly unstable conformations, a deterministic first-lower descent will, in most cases, converge to the same folding pathway taken by randomised descent. For three of the cases we considered, the number of observed local minima is slightly smaller for random descent in comparison to gradient descent. If

folding starts from highly unstable states e.g. cotranscriptional folding, then the question is how strongly differs the local minima ensemble produced by random descent in terms of structural features, free energy, and energy barriers when compared to deterministic descent.

Chapter 7

Approximating the number of local minima in partial RNA landscapes

The contents of this chapter appear in the following publication with additional unpublished results:

> [12] Andreas A. Albrecht, Luke Day, Ouala Abdelhadi Ep Souki and Kathleen Steinhöfel. "A new heuristic method for approximating the number of local minima in partial RNA energy landscapes", *Comp. Biol. & Chem.*, 60:43-52, 2016. doi: 10.1016/j.compbiolchem.2015.11.002.

7.1 Introduction

The analysis of energy landscapes plays an important role in mathematical modelling, simulation and optimisation. Among the main features of interest are the number and distribution of local minima within the energy landscape. The problem of calculating metastable RNA secondary structures (local minima) is considered, for example, in [98, 191, 196]. The `RNAsubopt` tool by Wuchty *et al.* [90] together with the `barriers` program [98] allows the user, in principle, to identify all metastable conformations within an energy range ΔE above the MFE conformation. However, due to the rapid increase of conformations with increasing ΔE, the approach is only applicable to short sequences or small values of ΔE.

Lorenz and Clote [191] describe the extension of computing the partition function over the set of locally optimal structures in RNA energy landscapes to

the `RNAlocopt` tool that computes the exact, total number of locally optimal structures and the exact partition function for locally optimal structures, along with the capability of identifying the set of local minima. The underlying energy model is the Turner nearest neighbour model [66] without dangles. The algorithm is an extension of McCaskill's algorithm [100], where locally optimal structures are accounted for by additional terms in the recursion scheme. Based upon dynamic programming, `RNAlocopt` computes the total number of locally optimal structures in $O(n^3)$ time, and the associated sampling of secondary structures takes $O(n^2)$ time, which is comparable to other tools. We note that in this work, we target partial RNA landscapes, i.e., the information about the total number of locally optimal structures for a given RNA sequence is not directly applicable. `RNAlocopt` has been recently modified to support Turner 2004 energy parameters. A detailed and sophisticated combinatorial analysis of the asymptotic behaviour of the number of saturated RNA secondary structures and locally optimal RNA secondary structures is carried out in [197] and [198] for a variety of energy models and structural constraints.

Saffarian *et al.* [196] present an algorithm for generating all locally optimal secondary structures assembled from a set of thermodynamically stable helices. The construction of locally optimal structures is divided into two steps: First juxtaposed base pair are processed, followed by nested (within juxtaposed positions) base pairs. The main step of the construction follows a recurrent relation which reminds of dynamic programming, although an estimation of worst case or expected run-time is not provided. Each element of the intermediate set of secondary structures is then further processed in order to generate locally optimal structures. The procedure is extended to the generation of locally optimal secondary structures from a given set of thermodynamically stable helices and computational experiments for six sequences of length up to 405 nt are presented.

Kucharík *et al.* [194] introduce basin hopping graphs as a new connectivity model of attraction basins within energy landscapes. Vertices represent local minima and edges connect vertices if the transition between the corresponding basins is energetically optimal in terms of the associated saddle point height. The authors present the two new tools `RNAlocmin` and `BHGbuilder` as basic implements for the approximation of basin hopping graphs. The tool `RNAlocmin` executes a modified Boltzmann sampling in order to generate sets of local minima. The modification of Boltzmann sampling tries to avoid oversampling of structures close to MFE conformation by using a parame-

terised thermodynamic temperature in Boltzmann weights. The `BHGbuilder` tool establishes the basin hopping graph by using a heuristic path-finding algorithm between the local minima identified by `RNAlocmin`. The authors present various comparisons to `RNAlocopt` from [191] regarding the coverage of local minima within a given time frame, which turn out to be in favour of `RNAlocmin`, partly with large differences in the number of detected local minima. However, one could argue that raising the temperature when running `RNAlocopt` could have increased the number of distinct locally optimal structures, which could affect the comparison to `RNAlocmin`. Kucharík *et al.* [194] estimate the applicability of the overall approach to a range of RNA sequence lengths bounded by about 300 nt.

7.1.1 Aims and contributions

Garnier and Kallel [199] proposed in 2002 a new sampling procedure for estimating the number of local minima. In this work, we focus on improved heuristic implementations of the general framework devised by Granier and Kallel with regard to run-time behaviour and accuracy of predictions. With aims at a fast and sufficiently accurate method for the evaluation of data obtained from partial RNA energy landscapes in the context of the approximation of the number of local minima. The partial energy landscapes can be defined by an energy offset above the minimum free energy conformation or by a bounded distance in terms of elementary transition steps from a given conformations within the energy landscape. As a continuation of work presented in [200], we consider energy landscapes induced by RNA secondary structures and partial energy landscapes defined by energy offsets. The method can be used, e.g., in a pre-processing step before starting a comprehensive analysis (or complete generation) of local minima in partial energy landscapes for *a priori* information about the expected number of local minima. Here, we focus on the fast and reliable evaluation of data produced by steepest decent that starts out from samples of secondary structures, whereas for pre-processing steps, such as initial sample size selection and randomised generation of sample sets, we rely on existing tools for RNA secondary structure generation and free energy calculation.

As in [200], we utilise the framework presented by Garnier and Kallel in [199] for approximating the number of local minima in a fitness landscape: At the initial stage, M elements of the landscape are randomly selected. Each

of the elements then 'moves' towards a local/global minimum S_m based upon steepest descent (being part of the attraction basin of S_m). This way, local minima 'collect' instances originating from the M initial landscape elements, and the number of local minima having 'collected' exactly j of the M elements is denoted by β_j. The method proposed in [199] tries to utilise the information about the distribution of β_j for a prediction about the total number of local minima within the landscape. The authors associate with the normalised sizes of attraction basins the Gamma distribution, where an approximation of the crucial parameter γ of the density function can be obtained by using (a) a basic equation established in [199] for linking the Gamma distribution to expected values $\beta_{j,\gamma}$ of sampling data β_j and (b) the χ^2-test with regard to $\beta_{j,\gamma}$ and β_j. The method has been applied in [200] to the approximation of the number of local minima in partial RNA folding landscapes. The application employs a minimisation procedure for the χ^2-test over a square grid for two parameters (γ, r), where γ defines the density function and r is an auxiliary parameter that eventually leads to the required approximation. Since the χ^2-minimisation is running over a square grid, finding suitable (γ, r) is relatively time-consuming. Moreover, as demonstrated in [200], the quality of approximations is affected by the values of $\beta_{j,\gamma}$ for large j (called tail values of $\beta_{j,\gamma}$, where j is close to the maximum j such that $\beta_j > 0$), together with large gaps between non-zero β_j for increasing j. The problem with tail values was observed and highlighted already in [199], see Section 5.3 therein. In this work, both problems, i.e., tail values and χ^2-minimisation, are addressed by a new heuristic that utilises a specific pooling procedure for tail values and substitutes the simultaneous χ^2-minimisation over a square grid by two linear χ^2-tests executed one after another and for γ-approximations only.

We aim at improved approximations of the number ν of local minima in partial RNA folding spaces. The approximation of ν can then be used for evaluating the outcome of procedures searching for local minima as presented in [194]. *A priori* knowledge about approximations of ν provides information about the distance of the current number of identified local minima to the true number of metastable conformations. In the present application, the number of instances as well as energy parameters of partial landscapes above minimum free energy conformations are chosen in such a way that a comparison to data generated by `RNAsubopt` and `barriers` is computationally feasible. `RNAsubopt` and `barriers` are used to calculate all secondary structures and the true set of local minima, respectively, within the partial energy

landscapes. The time complexity of our heuristic method can be estimated by $O\big(n^2 E_n D \max\{M,\nu\}\big)$, where M is the number of samples (initial secondary structures), E_n is the energy evaluation complexity (update of energy values when samples are 'moving' within the partial landscape), and D is the maximum number of steps executed in steepest descent (determined by the maximum energy difference between initial sample structures and their corresponding 'target' local minima). Our computational experiments suggest $M = O\big(\nu\big)$ with a relatively small constant factor, which implies a run-time estimation of $O\big(n^2 E_n D\nu\big)$. Since we are using `RNAlocmin` [194] with energy updates in neighbourhood transitions, we can assume $E_n = O\big(1\big)$, which leads to $O\big(n^2 D\nu\big)$. The evaluation procedure itself, which processes the β_j values obtained by steepest descent and is the main subject here, is relatively fast and terminates on standard PC desktop equipment in less than one second for all sample numbers M we consider here (including the test case $M = 30,000$).

7.2 RNA Sequences and partial landscapes

Firstly, we introduce the concept of energy landscapes in the context of folded RNA structures, including information about the sequences and data we are using in our computational experiments. Secondly, we explain the mathematical background of the stochastic method devised by Garnier and Kallel [199] for approximations of the number of local minima in fitness landscapes. Finally, we present the new method we propose for processing the data generated by the Garnier-Kallel algorithm.

7.2.1 Energy landscape definition

Our new approximation methods is demonstrated for the case of metastable conformations (local minima) of RNA secondary structures. In formal terms, the folded structure of an RNA sequence of length n is a node-labelled, undirected graph $G = (V,E)$, where $V = \{1,...,n\}$, $E \subseteq V \times V$ and $L(V) = \{A,C,G,U\}$, such that

(1) $(i,j) \in E \Leftrightarrow (j,i) \in E$;

(2) $\forall i\big(i \in \{1,...,n-1\} \rightarrow (i,i+1) \in E\big)$ (backbone bonds);

(3) For $1 \leq i \leq n$, there exists *at most* one $j \neq i, i \pm 1$, such that $(i,j) \in E$, where $L(i)$ and $L(j)$ comply with Watson-Crick pairs or G–U (U–G);

We note that(1)–(3) define tertiary structures, i.e. so-called pseudo-knots are allowed. The additional condition that

(4) $1 \leq i < k < j \leq n$, $(i,j) \in E$ and $(k,\ell) \in E$ imply $i \leq \ell \leq j$ defines so-called secondary structures (outer-planar graphs, where 'knots' are disallowed).

In this work, the basic structural properties of secondary structures as well as the associated energy parameters comply with standard settings as provided by the Vienna RNA package [65].

RNA folding landscapes

Given an RNA sequence R, we denote by $\mathsf{L}(R) = [\mathsf{C}, \mathsf{N}, E]$ the energy landscape defined by the set of secondary structures C, the neighbourhood relation N and the energy function $E : \mathsf{C} \to \mathbb{R}$. The conformation space C consists of secondary structures with standard settings as provided, for example, by the `RNAfold` tool [65], i.e. no isolated base pairs and at least three nucleotides in loops. Given a secondary structure S of sequence R, the neighbourhood N_S is defined by two types of single-step transitions $S \to S' \in \mathsf{N}_S$:

(1) Addition of one or two base pairs: a single base pair is added, if an existing helix is extended; two base pairs are added, if an unpaired position admits such an extension without extending a helix by two base pairs; the addition must ensure that the condition for the minimum loop size is not violated.

(2) Deletion of one or two base pairs: a single base pair is deleted as part of a helix, if at least two adjoined base pairs remain; otherwise, two base pairs are deleted.

The neighbourhood N_S covers all conformations that can be generated by a single application of one of the transitions, where by definition the secondary structure S itself belongs to N_S. It is important to note that a local/global minimum S_m is defined by $\forall S\big(S \in \mathsf{N}_S \to E(S) \geq E(S_m)\big)$. Thus, the case $\forall S\big(S \in \mathsf{N}_S \to E(S) = E(S_m)\big)$ is included and allows us to achieve a match between the set of local minima returned by `barriers` and the steepest descent procedure described below. The neighbourhood operations do not include the shift move as defined in [93] and accounted for in [200]. The reason lies in the

newer version of `RNAsubopt` (2.x), and the analysis of the `barriers` source code reveals that the shift move is not part of the implementation for secondary structures where isolated base pairs are not admitted. This also results in different values of the number of local minima for the sequences considered in [200]. However, for consistency with `RNAsubopt` and `barriers`, we employ `RNAlocmin` [194] for executing steepest descent.

The single-step transitions (1) and (2) are utilised in a steepest descent algorithm that is executed within $\mathsf{L}(R)$. The steepest descent algorithm can be described as follows:

(i) Initialise $S_0 = S$.

(ii) For $u \geq 0$, set $S_{u+1} = \texttt{argmin}_{S' \in \mathsf{N}_{S_u}} E(S')$.

(iii) If $E(S_{u+1}) < E(S_u)$, then $u = u+1$ and `goto` (ii), otherwise terminate with S_u.

Step (ii) implicitly assumes that there exists only a single $S' = S_{u+1} \in \mathsf{N}_{S_u}$ that minimises the function $E(S)$. In general, there is no guarantee that this holds for energy landscapes induced by RNA secondary structures. However, in our computational experiments we regularly checked the condition for a large number of steepest descent procedures and no violation was observed.

If the function $E : \mathsf{L}(R) \to \mathbb{R}$ always produces a single minimum in step (ii), then the steepest descent procedure is deterministic and leads to a partition of $\mathsf{L}(R)$ into attraction basins A_v of local/global minima m_v, $v = 1, \ldots, \nu$, where $A_u \cap A_v = \emptyset$ for $u \neq v$, $\bigcup_{v=1}^{\nu} A_v = \mathsf{L}(R)$, and ν is the total number of local minima within $\mathsf{L}(R)$, including the global minima. The steepest descent procedure terminates at the local minimum that defines the attractions basin A_v, i.e. A_v consists of all S for which (ii) and (iii) lead to the minimum m_v, $v = 1, \ldots, \nu$.

3' UTR sequences

Out of the sequences studied in [200], we selected five longer sequences for the application-specific fine-tuning of the new heuristic method for evaluating the β_j data. Additionally, four sequences from [9] plus one sequence of length 504 nt were selected for independently testing the approach. The selection of sequences is governed by the need of having verifiable data about the set of conformations in partial landscapes, including information about the number and structure of local/global minima. For each sequence, the set

of conformations is generated by the `RNAsubopt` program (version 2.0.7) [90], and the number and structure of local/global minima is provided by the `barriers` implementation (version 1.5.2) [65]. For an offset ΔE above the minimum free energy conformation (global minimum), `RNAsubopt` produces a set of conformations that exponentially increases depending on ΔE. In order to be able to execute a large number of computational experiments on each of the sequences, we selected as partial landscapes the subset of conformations within the ΔE range above the minimum free energy conformation. The selection allows us to study sequences of length up to about $\ell = 500$ nt for a wide range of sample sets $\mathsf{M} = \{S_1, S_2, ..., S_M\}$.

The ten sequences represent 3'UTRs of human RNAs, where the information about the sequences is drawn from the NCBI Nucleotides database and the Ensembl database. More details about the sequences are given in 7.1. The Transcript IDs are from the Ensembl database.

No	Gene Name	ℓ	NCBI Ref. No.	Transcript ID
1	MRPL9	407	NM_031420.2	ENST00000368830
2	ALDH4A1	400	NM_170726.2	ENST00000290597
3	GMEB1	113	NM_024482.2	ENST00000361872
4	PAX7	99	NM_002584.2	ENST00000375375
5	OXT	99	NM_000915.3	ENST00000217386
6	AQP5	504	NM_001651.3	ENST00000293599
7	HLA-G	386	NM_002127.5	ENST00000360323
8	HTR3E	302	NM_001256614.1	ENST00000360323
9	CBR1	284	NM_001757.2	ENST00000290349
10	LIG3	124	NM_002311.4	ENST00000262327

Table 7.1 3' UTR Sequences, ℓ denotes the length of sequences (number of nucleotides).

The selection of sequences is partly based on the results of microRNA target predictions, with the main aim of collecting sequences of a certain length. For example, TargetScan (version 6.2) [120] predicts NM_031420.2 as a conserved target of *hsa-miR-21* and *hsa-miR-590-5p*, which is also supported by MicroCosm (miRanda) [105]. NM_170726.2 is predicted as a conserved target of *hsa-miR-184*, and the same miRNA is also predicted by MicroCosm. Among the five test sequences, the 3' UTRs HLA-G, HTR3E, CBR1 and LIG3 are related to microRNA target prediction in the context of single nucleotide polymorphisms [9]. Sequence *R*6 (NM_001651.3) is predicted to be a target of

hsa-miR-96. Furthermore, we tried to design a set of sequences with varying ratios of the size of the subspace vs the number of local minima within this subspace, including sequences with 'rugged' partial landscapes close to the minimum free energy conformation.

7.2.2 Partial energy landscapes

Table 7.2 displays information about the subsets $\mathsf{C}_{\Delta E} \subset \mathsf{C}(R)$ (partial energy landscapes) associated with each of the sequences R from Table 7.1 and the energy range ΔE above the minimum free energy conformation. Since we are using the updated versions of `RNAsubopt` and `barriers` (no shift operation in neighbourhood transitions), the data reported in 7.2 for the first five sequences differ from the corresponding values presented in [200].

No	R	ℓ	ΔE	$\lvert\mathsf{C}_{\Delta E}\rvert$	ν	$\lvert\mathsf{C}_{\Delta E}\rvert/\nu$
1	NM_031420.2	407	3.7	126,906	1,870	67.9
2	NM_170726.2	400	2.4	18,569	2,020	9.2
3	NM_024482.2	113	4.5	12,381	2,322	5.3
4	NM_002584.2	99	9.0	84,725	2,401	35.3
5	NM_000915.3	99	7.0	24,609	1,440	17.1
6	NM_001651.3	504	2.6	12,518	3,054	4.1
7	NM_002127.5	386	1.6	9,609	2,982	3.2
8	NM_001256614.1	302	4.2	17,371	2,591	6.7
9	NM_001757.2	284	3.1	16,580	3,272	5.1
10	NM_002311.4	124	6.0	20,646	3,441	6.0

Table 7.2 Sequences and associated partial energy landscapes.

The sequences exhibit varying values of the ratio $\mathrm{ru}(\Delta E, R) = |\mathsf{C}_{\Delta E}|/\nu$, which can be seen as a simplified measure for the 'ruggedness' of energy landscapes. In the strong sense, rugged energy landscapes (e.g., in protein folding simulations) are associated with many local minima and high energy barriers separating local minima, cf. [201]. However, in this work we are dealing with steepest descent within partial energy landscapes (and not with overcoming high energy barriers), and therefore we think that using $\mathrm{ru}(\Delta E, R)$ as a measure for 'ruggedness' is justified, since the value of $\mathrm{ru}(\Delta E, R)$ affects the relation between the size of M and $\mathrm{lm}(\mathsf{M})$. The smallest value of $\mathrm{ru}(\Delta E, R)$ (i.e., highest degree of 'ruggedness') is produced by NM_002127 (see Table 7.2). We note that for sequences No 1-5 the values of $\mathrm{ru}(\Delta E, R)$ can be ordered

in such a way that the successor is about twice the value of its predecessor, with the maximum value for sequence No 1. For sequences No 6-10, the maximum value of $|\mathsf{C}_{\Delta E}|$ is about twice the value of the minimum $|\mathsf{C}_{\Delta E}|$.

It is important to note that we achieved better approximation results when the secondary structures of a sample set M of size M were drawn from the top energy range of the given subspace $\mathsf{C}_{\Delta E}$ of conformations (`RNAsubopt` returns an enumerated list ordered by the free energy, which allows for a random selection of positions above a given free energy value. Depending upon the relation between the size $|\mathsf{C}_{\Delta E}|$ and M, the elements S of M were randomly selected from the top third or top quarter of $\mathsf{C}_{\Delta E}$ with respect to $E(S)$, if $|\mathsf{C}_{\Delta E}| >> 3M$.

7.2.3 Garnier-Kallel method

In this section, we follow the description of the stochastic approximation procedure devised by Garnier and Kallel as presented in [199] (specifically, in Section 5). Here, we consider the approximation problem called 'inverse problem' by the authors, whereas the 'direct problem' relates to the probability that for a given sample size M each attraction basin consists at least one conformation from the sample set.

Garnier and Kallel [199] introduce the normalised size $\alpha_v = |A_v|/|\mathsf{C}|$ for attraction basins A_v (see Definition 2.1 in [199]) and assume a parameterised random distribution of $z = \alpha_v$ with the density function p_γ defined by

$$p_\gamma(z) = \frac{\gamma^\gamma}{\Gamma(\gamma)} \frac{z^{\gamma-1}}{e^{\gamma z}}, \tag{7.1}$$

where $\gamma > 0$ and $\Gamma(x) = \int_0^\infty e^{-y} y^{x-1} dy$ is the Euler function. The main task is to approximate the density function p_γ by a sampling method over attraction basins, which eventually leads to an approximation of the value of γ.

Let $\mathsf{M} = \{S_1, S_2, ..., S_M\} \subset \mathsf{C}$ denote a sample set of randomly selected secondary structures. For each of the S_u, the steepest descent procedure is executed. The application of neighbourhood transitions (1) and (2) transforms an individual conformation S_u into a local minimum m_v, if S_u belongs to the attraction basin A_v associated with m_v. By $\mathsf{D}(m_v) = \mathsf{M} \cap A_v$ we denote the set of conformations from M such that steepest descent terminates at m_v, $v = 1, ..., \nu$. Therefore, under the assumption that the attraction basins of all ν local/global minima generate a partition of C, the application of the steepest

descent algorithm to the elements of M identifies an individual set of local minima. The set of local minima induced by M is denoted by

$$\mathrm{LM}(\mathsf{M}) = \{m | \text{ size of } \mathrm{D}(m) > 0\}, \tag{7.2}$$

and we set $\mathrm{lm}(\mathsf{M}) = |\mathrm{LM}(\mathsf{M})|$.

A drawback of our current approach and parameter settings (energy value calculations) is the fact that local minima S_m may exhibit a neighbourhood with the property $\forall S\big(S \in \mathsf{N}_S \rightarrow E(S) = E(S_m)\big)$. Therefore, different local minima might be located within the same attraction basin. Merging such local minima into a meta-conformation and re-calculating the results of steepest could resolve the problem, but we noticed that the approximation results are affected only marginally. Consequently, local minima from the same attraction basin are counted separately.

Let $B_j = \{\mathrm{D}(m) | \text{ size of } \mathrm{D}(m) = j\}$ denote the set of sets $\mathrm{D}(m)$ that have the same size j. Following the notations provided in [199], we then set

$$\beta_j = |B_j|, \; j \geq 0. \tag{7.3}$$

By this definition, β_0 is the number of local minima not 'visited' by any of the $M = |\mathsf{M}|$ steepest descent procedures. Since β_j minima are 'visited' by j elements of M via steepest descent, we have

$$M = \sum_{j=0}^{M} j\beta_j. \tag{7.4}$$

Furthermore, (7.2) defines the set of local minima 'visited' by at least one element of M. Each of such local minima is counted exactly by one set B_j, and therefore we obtain

$$\mathrm{lm}(\mathsf{M}) = \sum_{j=1}^{M} \beta_j, \tag{7.5}$$

which, in other terms, is the number of observed local minima. Figure 7.1 illustrates Eqn. 7.4 and Eqn. 7.5 for the case of $\nu = 5$ and $M = 8$, i.e. the landscape consists of five attractions basins.

We note that information about the values of β_j can be obtained from computational experiments executed over (subsets of) C or, more practically, over the set of conformations from a pre-defined N_S^k. The problem now is to relate such observed values to independently calculated approximations

of β_j. Let $\beta_{j,\gamma}$ denote the expected value of β_j under the assumption that the normalised sizes α of attraction basins A are distributed according to p_γ defined in (7.1), i.e. $\beta_{j,\gamma} = \mathbb{E}_\gamma[\beta_j]$. Garnier and Kallel [199] established the relation

$$\beta_{j,\gamma} = \nu \binom{M}{j} \frac{\Gamma(\gamma+j)}{\Gamma(\gamma)} \frac{\Gamma(\nu\gamma)}{\Gamma((\nu-1)\gamma)} \frac{\Gamma((\nu-1)\gamma+M-j)}{\Gamma(\nu\gamma+M)}, \tag{7.6}$$

where $j = 1,\ldots,M$. In Eqn. 7.6, the value of ν is unknown. However, for $\nu = M/r$, $r > 0$, a fixed value of M, and appropriate approximations of the Γ-function, the $\beta_{j,\gamma}$ can be approximated according to Eqn. 7.6 as functions of (M, j, γ, r). Such a parameterised representation enables us to connect the computed values $\beta_{j,\gamma} = \beta_{j,\gamma}(M,r)$ to the observed values β_j.

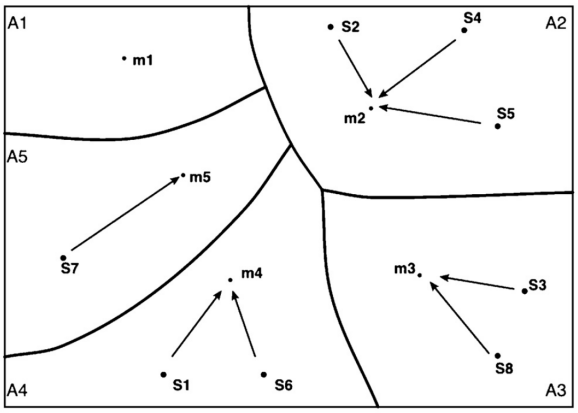

Fig. 7.1 Partition of energy landscape according to steepest descent. The values of $\mathrm{D}(m_v) = \mathsf{M} \cap A_v$ are: $\mathrm{D}(m_1) = \emptyset$, $|\mathrm{D}(m_2)| = 3$, $|\mathrm{D}(m_3)| = |\mathrm{D}(m_4)| = 2$, $|\mathrm{D}(m_5)| = 1$. Therefore, $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 2$, and $\beta_3 = 1$. Eqn. 7.4 is represented by $0\cdot 1+1\cdot 1+1\cdot 0+2\cdot 2+3\cdot 1 = 8$, which is the number M of sample structures. For Eqn. 7.5 we have $0+1+2+1 = 4 = \mathrm{lm}(\mathsf{M})$, i.e. four local minima are detected.

Due to the extremely large range of function values of $\Gamma(x)$ for settings of x as given in (7.6), we adopt the following approach of approximating the values of $\beta_{j,\gamma}$: We represent Eqn. 7.6 by

$$\beta_{j,\gamma}(M,r) = \frac{M}{r} \binom{M}{j} \frac{A_1}{A_2} \frac{B_1}{B_2} \frac{C_1}{C_2} \tag{7.7}$$

and then calculate the value

$$P = \ln \binom{M}{j} + \ln A_1 + \ln B_1 + \ln C_1 - \ln A_2 - \ln B_2 - \ln C_2. \tag{7.8}$$

The $\ln\Gamma(x)$ are determined by using a standard C^{++} function for $\Gamma(x)$, and for the binomial coefficient we use the formula

$$\ln\binom{M}{j} = \sum_{s=M-j+1}^{M} \ln s - \sum_{t=1}^{j} \ln t. \tag{7.9}$$

Finally, we set

$$\beta_{j,\gamma}(M,r) = \frac{M}{r} e^{P}. \tag{7.10}$$

Given the sequence of pairs $[\beta_{j,\gamma};\beta_j]$, $j = 1,.....,M$, of calculated values $\beta_{j,\gamma}$ and observed values β_j, the task is to identify a value for γ that provides a best fit of the $\beta_{j,\gamma}$ to the values of β_j. Garnier and Kallel [199] propose the χ^2-test for approximating γ ($j_{\max}$ is maximum j such that $\beta_j > 0$):

$$\min_{\gamma>0} T_\gamma \text{ for } T_\gamma = \sum_{j=1}^{j_{\max}} \frac{(\beta_j - \beta_{j,\gamma})^2}{\beta_{j,\gamma}}. \tag{7.11}$$

As shown in (7.7), we have chosen the parameterised representation $\beta_{j,\gamma} = \beta_{j,\gamma}(M,r)$, which leads to $\min_{\gamma,r>0} T_\gamma(r)$. In [200], we followed (7.11) and subsequently executed $\min_{\gamma,r>0} T_\gamma(r)$ for identifying a pair $[\gamma_a;r_a]$ that (approximately) minimises $T_\gamma(r)$. The approximation of the number of local minima can then be finalised by setting

$$\nu_a = \frac{M}{r_a}. \tag{7.12}$$

In the study [200], the pair $[\gamma_a;r_a]$ is determined by searching for a global minimum within a grid of size $[(r_{\max} - r_{\min})/\delta_r] \times [(\gamma_{\max} - \gamma_{\min})/\delta_\gamma]$, where δ_r and δ_γ define the corresponding elementary step-sizes. The lower bound $\gamma_{\min}$ of the γ-range was chosen equal to $\gamma_{\min} = 0.1$ in order to ensure the numeric stability of $\beta_{j,\gamma}(M,r)$-calculations. We achieved an average deviation of ν_a from values ν of about 34% over all eleven sequences considered in [200]. Furthermore, for all eleven instances the best approximations were obtained for values of γ_a close or equal to $\gamma_{\min}$. Therefore, we were searching for alternative ways of finding close approximation of β_j by $\beta_{j,\gamma}$, which is the main contribution.

Apart from a pooling procedure introduced for tail values of $\beta_{j,\gamma}(M,r)$, the new heuristic takes advantage of Eqn. 5.8 from [199]:

$$\frac{\sum_{j=1}^{j_{\max}} \beta_j}{M} = \frac{1 - \left(1 + \frac{r}{\gamma_0}\right)^{-\gamma_0}}{r} \tag{7.13}$$

where γ_0 is assumed to be known and r is identified from the equation. In our heuristic, we search within a given r-range for a value of r that minimises the absolute value of the difference between the LHS and RHS of (7.13). The new method is generic and resulted in a much lower average deviation from values ν. The method was devised on a subset of five sequences considered in [200] and then tested on five sequences related to miRNA target prediction that were partly analysed in [9] (see Chapter 4).

7.2.4 The main algorithm

The new method of evaluating the set of data β_j, $j = 1,...,M$, defined in (7.3) consists of five major steps, where four of the steps produce either a γ-approximation (γ_a, see Eq. 7.11) or an r-approximation (r_a):

(a) Identifying γ_a^1 from $p(z = \frac{j}{M}) \approx \beta_j \frac{j}{M}$ based upon (1) and χ^2-test over $\gamma \in [\gamma_{\min}; \gamma_{\max}]$.

(b) Identifying r_a^1 by using (7.13) for $\gamma_0 = \gamma_a^1$ and search over all $r \in [r_{\min}; r_{\max}]$.

(c) Identifying γ_a^2 by using (7.6) and (7.11) for $r = r_a^1$ and by utilising a new pooling procedure for tail values of $\beta_{j,\gamma}(M,r)$.

(d) Identifying $r_a = r_a^2$ by using again (7.13) for $\gamma_0 = \gamma_a^2$ and search over r as in (b).

(e) Selecting approximation ν_a of ν by setting $\nu_a = Z/r_a$, where Z is defined by the outcome of the pooling procedure used in (c) and related to M and the number of observed local minima $\mathrm{lm}(\mathsf{M})$; see Eqn. 7.5.

In more detail, we proceed as follows: We assume that steepest descent applied to the sample set $\mathsf{M} = \{S_1, S_2, ..., S_M\}$ produces the values β_j, $j = 1,...,M$, as defined in (7.3). The data are assumed to be generated by the l^{th} trial of randomly generated sample sets, i.e. $\mathsf{M}_l = \mathsf{M}$ and $|\mathsf{M}_i| < |\mathsf{M}_l|$ for sample sets M_i from preceding steps $1 \leq i < l$. The evaluation of the data β_j then consists of the following five major steps:

(A) Calculating γ_a^1: Since the $|A_v|$-distribution is not known *a priori* in a black-box-scenario, the distribution is approximated by a limited number M of samples given by M. We recall that β_j is the number of local minima that 'attract' j out of the M samples. Thus, in our heuristic

approach we assume that the probability p_γ, as defined in (7.1), of having a normalised size j/M of the attraction basin is approximated by $\beta_j(j/M)$. We select a γ-range $\gamma_{\min} \leq \gamma \leq \gamma_{\max}$ and a step-size $\delta_\gamma = 0.01$, i.e. $\gamma_{n+1} = \gamma_n + \delta_\gamma$, where for reasons of numeric stability $\gamma_{\min} = 0.25$. Based upon the shape of $p_\gamma(z)$ (see, e.g., Figure 3 in [199]) and the typical sequence of values of β_j, we select $\gamma_{\max} = 2.25$. Consequently, we are searching within $[\gamma_{\min}; \gamma_{\max}]$ for an approximate solution γ_{a}^1 of

$$\min_\gamma \sum_{j=1}^{j_{\max}} \frac{(p_\gamma(j/M) - (j\beta_j)/M)^2}{p_\gamma(j/M)} \,|_{\beta_j>0}\,. \tag{7.14}$$

We note that the positions j with $\beta_j > 0$ are indexed, which contributes to a significant speed-up and is utilised throughout the approximation procedure.

(B) Calculating r_{a}^1: We select the step-size $\delta_r = 0.01$ and a lower bound $r_{\min}$, i.e. $r_{i+1} = r_i + \delta_r$. Taking into account (7.12), the lower bound $r_{\min}$, obviously, limits the 'lookahead' of our approach. We incorporate the knowledge about the number of observed local minima $\mathrm{lm}(\mathsf{M})$ and set $r_{\min} = (M - \mathrm{lm}(\mathsf{M}))/M$. Since $\mathrm{lm}(\mathsf{M}) \leq \nu$, Eqn. 7.12 justifies the setting $r_{\max} = M/\mathrm{lm}(\mathsf{M})$. We are then searching within $[r_{\min}; r_{\max}]$ for an approximate solution r_{a}^1 of

$$\min_r \Big| \frac{\sum_{j=1}^{j_{\max}} \beta_j}{M} - \frac{1 - \left(1 + \frac{r}{\gamma_{\mathrm{a}}^1}\right)^{-\gamma_{\mathrm{a}}^1}}{r} \Big|, \tag{7.15}$$

see also Figure 7.2. Thus, (A) and (B) produce an initial pair $(\gamma_{\mathrm{a}}^1; r_{\mathrm{a}}^1)$ of crucial parameters by avoiding minimisation over a square grid as in [200].

(C) Calculating γ_{a}^2 with pooling of $\beta_{j,\gamma}$ tail values: In our approach, we apply the χ^2-test from (7.11) to $\beta_{j,\gamma}$ values defined in (7.6) and (7.7). Properties, applications and limitations of the χ^2-test have been discussed for a long time, see [202]. Among the features often highlighted as limiting the applicability are small expected values (in our case, $\beta_{j,\gamma}$). Frequent recommendations expressed in the literature of how to deal with small expected values mainly comprise of two rules: (a) a lower bound $\beta_{j,\gamma} \geq h = 5$, i.e. the summation in (7.11) includes only positions

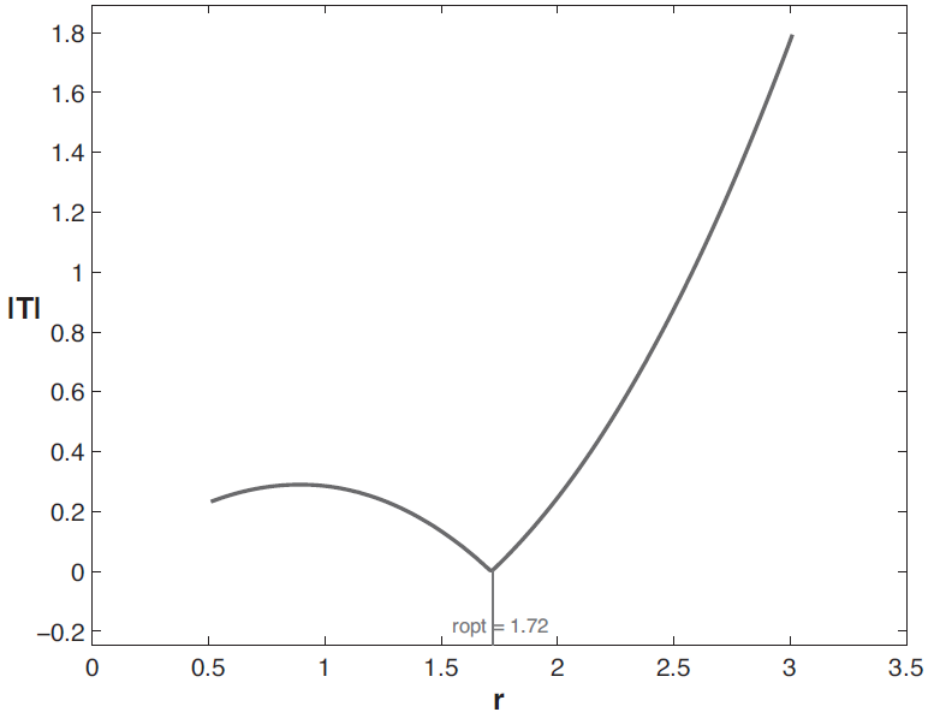

Fig. 7.2 Finding r_a via minimising the absolute value of $T = \text{LHS} - \text{RHS}$ in Eqn. 7.13. Represented is $|T|$ for NM_024482.2 and $M = 4,000$ as a function of r.

j that meet the condition; (b) combining $\beta_{j,\gamma} < h$ in successive positions into a single position until the sum is equal to or exceeds the lower bound h, which is often called 'pooling.'

In the present application, the expected values $\beta_{j,\gamma}$ relate to the number $|B_j|$ of local minima such that for j sample structures steepest descent terminates in an individual local minimum from B_j. The case of single local minima attracting a large number of sample structures can be frequently observed in partial energy landscapes induced by RNA secondary structures, i.e., in other terms, $\beta_{j,\gamma} \approx 1$ is a legitimate setting for large j. Therefore, we selected $h = 1$ instead of $h = 5$. The selection is supported by computational experiments on sequences from 7.2, with better approximations for $h = 1$.

We apply pooling with respect to $h = 1$, where we do not introduce new notations for the modified β_j and $\beta_{j,\gamma}$. It is important to note that the $\beta_{j,\gamma}$-values from (7.7) are calculated for the fixed r_a^1 and the pooling procedure is executed for varying $\gamma = \gamma_{n+1} = \gamma_n + \delta_\gamma$ from $[\gamma_{\min}; \gamma_{\max}]$:

- ◇ If j is the smallest number such that $\beta_{j-1,\gamma} \geq 1$ and $\beta_{j,\gamma} < 1$, we initialise $s = 0$, $X_j = \beta_{j+s}$, and $Y_j = \beta_{j+s,\gamma}$.
- ◇ For $s = s+1$, we calculate $X_j = X_j + \beta_{j+s}$ and $Y_j = Y_j + \beta_{j+s,\gamma}$, until for the first time $Y_j \geq 1$ or all $\beta_j > 0$ are covered by reaching $j_{\max}$.
- ◇ We set the modified values $\beta_j = X_j$ and $\beta_{j,\gamma} = Y_j$, and repeat the procedure, starting with $j = j+s+1$, until all $\beta_j > 0$ are covered at $j_{\max}$. Thus, for the modified $\beta_{j,\gamma}$ with the largest index j (denoted by j_{m}) we have, in general, $\beta_{j_{\mathrm{m}},\gamma} < 1$.

For the modified β_j and $\beta_{j,\gamma}$, the χ^2-test from (7.11) now turns into minimising the value of

$$T(\gamma) = \sum_{j=1}^{j_{\mathrm{m}}-1} \frac{(\beta_j - \beta_{j,\gamma})^2}{\beta_{j,\gamma}}, \tag{7.16}$$

where we assume $\beta_{j_{\mathrm{m}},\gamma} < 1$ (other cases not observed). Based on the new values β_j, we now define

$$\widetilde{M} = \sum_{j=1}^{j_{\mathrm{m}}-1} j\beta_j. \tag{7.17}$$

Furthermore, we denote by $\widehat{M}$ the number of initial samples counted by β_j from (7.3) where the pooling procedure was not applied since $\beta_{j,\gamma} \geq 1$ (corresponding to 'unpooled' positions j). We note that both $\widetilde{M}$ and $\widehat{M}$ depend on the particular γ under consideration (index omitted). Furthermore, j_{up} denotes the largest j such that $\beta_{j,\gamma} \geq 1$ and $\beta_{j+1,\gamma} < 1$ (denotes the number of 'unpooled' positions; here, $\beta_{j,\gamma}$ is from (7.7) and not the result of pooling).

Now, for each $\gamma \in [\gamma_{\min}; \gamma_{\max}]$, the $\beta_{j,\gamma}$ are calculated according to (7.7), the pooling procedure is applied to β_j and $\beta_{j,\gamma}$, and the value of $T(\gamma)$ is calculated according to (7.16). The γ that minimises $T(\gamma)$ defines the value of γ_{a}^2.

(D) Calculating r_{a}^2: The same procedure as in (B) is applied, however, for (a) $\gamma_0 = \gamma_{\mathrm{a}}^2$ in (7.13) and (b) for $\sum_{j=1}^{j_{\mathrm{m}}-1} \beta_j/\widetilde{M}$ in (7.13), where the β_j are from the pooling procedure executed for $(\gamma_{\mathrm{a}}^2; r_{\mathrm{a}}^1)$. The result is the final r-approximation $r_{\mathrm{a}} = r_{\mathrm{a}}^2$.

(E) Calculating ν_a: As in (7.12), the approximation ν_a of the true number ν of local minima is calculated according to

$$\nu_a = \frac{Z}{r_a}, \tag{7.18}$$

where Z is determined by a small set of rules. We observed that the value of M reduced by the number $\mathrm{lm}(\mathsf{M})$ of observed local minima is a useful parameter for the selection of Z; i.e., we set $M_{\text{lm-r}} = M - \mathrm{lm}(\mathsf{M})$. We use the relation of $M_{\text{lm-r}}$ to $M \geq \widetilde{M} \geq \widehat{M}$ for determining the value of Z, where for relatively large values of $\mathrm{lm}(\mathsf{M})$ with $M_{\text{lm-r}} < \widehat{M}$ (number of samples covered by 'unpooled' positions j) we further distinguish between cases where $\widehat{M} - M_{\text{lm-r}}$ is smaller or larger than $\widetilde{M} - \widehat{M}$. Thus, we have four major cases defined by $M_{\text{lm-r}} > \widetilde{M}$, $\widetilde{M} \geq M_{\text{lm-r}} \geq \widehat{M}$, $\widehat{M} - M_{\text{lm-r}} > \widetilde{M} - \widehat{M} \geq 0$, and $0 < \widehat{M} - M_{\text{lm-r}} \geq \widetilde{M} - \widehat{M}$. For each of the four major relations, two subcases are considered that depend, e.g., on the value of $M - \beta_{j_m}$ in relation to $\widetilde{M}$ and $\widehat{M}$ (many samples can be covered by β_{j_m}, sometimes more than $\mathrm{lm}(\mathsf{M})$, which reflects the existence of a few local minima with large attraction basins). For the eight (sub-) cases, the corresponding values of Z are of the type M, $M - \beta_{j_m}$, $(M + \widetilde{M})/2$, $(M + \widehat{M})/2$, $(\widehat{M} + M_{\text{lm-r}})/2$, and $\widehat{M} + \beta_{j_m}$. Here, $Z = M$ covers the two cases where $M_{\text{lm-r}} > \widetilde{M}$ with only a single pooled position from (D) and $\widetilde{M} \geq M_{\text{lm-r}} \geq \widehat{M}$ with at least two pooled positions from (D), and $Z = (\widehat{M} + M_{\text{lm-r}})/2$ covers the two cases $M_{\text{lm-r}} > \widetilde{M}$ with at least two pooled positions from (D) and $M_{\text{lm-r}} < \widehat{M}$ together with $\widehat{M} - M_{\text{lm-r}} > \widetilde{M} - \widehat{M}$ and a single pooled position from (D). Additionally, we separate the subcase $j_{up} = j_m - 1$ (only a few local minima with 'moderate' size of attraction basin), where, however, values Z from the same range are used, and the subcase was observed only for the two major cases determined by $M_{\text{lm-r}} < \widehat{M}$.

Thus, from the five instances used for the design of the method, $8+2=10$ subcases were identified by running each of the instances for about six different values of M.

7.3 Results

The core data presented are from 33 runs defined by different values of M and the partial energy landscapes as displayed in Table 7.2 for each of the sequences. For a given sequence, each run takes as input the value of M, the set of data β_j, j = 1, ..., M, and - for comparison purposes - the value of ν. The step sizes δ_γ and δ_r are fixed to $\delta_\gamma = \delta_r = 0.01$, and the search range is $[\gamma_{\min}, \gamma_{\max}] = [0.25, 2.25]$ by default and for $[r_{\min}, r_{\max}]$ calculated individually as explained in (B). The data β_j as well as the corresponding ν are obtained in a pre-processing step by steepest descent as described in Section 7.2.1.

Since in the present study the size of the partial energy landscape is known from `RNAsubopt`, the initial setting M_1 can be selected in the following way: In [191], the authors argue that $\nu \sim \sqrt{|\mathsf{C}_{\Delta E}|}$ (a rough estimation of $|\mathsf{C}_{\Delta E}|$ is sufficient, if applied to a neighbourhood region of fixed maximum distance to a given sequence). While the value of $\sqrt{|\mathsf{C}_{\Delta E}|}$ indeed can be taken as the initial sample size, we found that rounding up to the next order of magnitude can speed up the approximation procedure. For example, for NM_170726 (ALDH4A1) we have $\sqrt{|\mathsf{C}_{2.4}|} \approx 136$, and consequently one can select $M_1 = 1,500$ for the first approximation. For M_2, one can select $M_2 = M_1 + \Delta M$ with ΔM in the region of $\max\{\mathrm{lm}(\mathsf{M}_1); M_1/10\}$.

After approximations in accordance with (7.18) are obtained for M_1 and M_2, useful auxiliary data can be extracted from the execution of (A)-(E). Of particular interest is the gain in observed local minima $\mathrm{lm}(\mathsf{M})$ when moving from M_1 to M_2. For two (subsequent) values of $M_a < M_b$ and associated $\mathrm{lm}(\mathsf{M})$ we calculate:

$$\mathrm{gain}_{\mathrm{frac}}(M_b) = \left|\frac{M_a}{M_b - M_a} \cdot \frac{\mathrm{lm}(\mathsf{M}_b) - \mathrm{lm}(\mathsf{M}_a)}{\mathrm{lm}(\mathsf{M}_a)}\right|. \tag{7.19}$$

The value of $\mathrm{gain}_{\mathrm{frac}}$ can be utilised for the decision about the termination of subsequent increases of values of M.

Table 7.3 shows the results for three different values of M for each of the sequences and parameter settings as given in Table 7.2. By Δ_{a} we denote the percentage of the deviation of ν_{a} from ν (percentage relative to ν; the notation is in line with Eqn. 7.18). Accordingly, Δ_{sel} is equal to Δ_{a} for the sample size M selected for the analysis of best approximations.

For HLA-G ($\ell = 386$) we executed a very long run with $M = 30,000$ over a partial landscape with the higher value of $\delta E = 2.4$ kcal/mol (in 7.2, we have

$\Delta E = 1.6$kcal/mol). The number of local minima returned by `barriers` is 22,662, hence $M/\nu = 1.32$. The critical procedures (A) – (D) proved to be numerically stable. For the ratio $M/\nu = 1.32$, the results were in the region of expected values, with $\Delta_a = 8.48\%$ and a run-time of the β_j evaluation (procedures (A) – (D)) below 0.5 sec.

The sub-procedure (E) described in Section 7.2.4 utilises the value of $M_{\text{lm-r}} = M - \text{lm}(\mathsf{M})$. Since $\text{lm}(\mathsf{M}) \leq \nu$, we observe for the same sequence but increasing M changing selection cases out of the 10 (sub-)cases. For example, $M'_{\text{lm-r}} < \widetilde{M'}$ can change at the next step with $M'' > M'$ to $M''_{\text{lm-r}} > \widetilde{M''}$. This may lead to an intermediate increase of Δ_a, as, for example, for sequences No. 6 and No. 8.

Regarding the termination of subsequent increases of M for a particular RNA sequence (partial folding landscape), we distinguish between three cases:

(I) For increasing M, only the selection case $M_{\text{lm-r}} > \widetilde{M}$ applies;

(II) For increasing M, one of the selection cases with $M_{\text{lm-r}} \leq \widetilde{M}$ or $M_{\text{lm-r}} < \widehat{M}$ applies, the latter together with $\widehat{M} - M_{\text{lm-r}}$ smaller than $\widetilde{M} - \widehat{M}$;

(III) For increasing M, only $\widehat{M} - M_{\text{lm-r}}$ larger than $\widetilde{M} - \widehat{M}$ applies.

Case (I) applies to sequences No. 1 and No. 4; we note that for both sequences the number of observed local minima $\text{lm}(\mathsf{M})$ is small relative to M. Case (III) applies to sequences No. 6 and No. 7, where for increasing M the number of observed local minima increases steadily. Based on the data presented in Table 7.3, we suggest the termination with M for Case (I), if $\text{gain}_{\text{frac}}(M) \lesssim 1/3$; for Case (II), if $\text{gain}_{\text{frac}}(M)$ is for two subsequent M close to or below 1/2; for Case (III), if $M/\nu_a \gtrapprox 1.5$.

We recall that the M samples are randomly selected from the top energy range of $\mathsf{C}_{\Delta E}$ in order to capture many local minima. If the M samples are drawn from the entire set $\mathsf{C}_{\Delta E}$, the number of observed local minima $\text{lm}(\mathsf{M})$ can be significantly smaller. For example, for sequence No 1 (MRPL9; largest value of $\text{ru}(\Delta E, R)$, see 7.2), we have for a run with $M = 6,000$ and random selection over the entire range only 683 local minima vs 826 recorded in Table 7.3 (21% more); for sequence No 4 (PAX7; second largest value of $\text{ru}(\delta E, R)$) and $M = 6,000$, the corresponding values are 946 vs 1020 (8% more); for sequence No 3 (GMEB1; smallest value of $\text{ru}(\delta E, R)$) and $M = 3,500$, the corresponding values are 1175 vs 1375 (17% more). This affects also the

No	R	M	lm(M)	$\text{gain}_{\text{frac}}$	r_{a}	v_{a}	$\Delta_{\text{a}}\%$	$\Delta_{\text{sel}}\%$
1	$\ell = 407$	5000	754	0.36	3.45	1450	22.46	
	$v = 1870$	6000	826	0.48	3.76	1595	14.71	
		7000	865	0.28	4.13	1696	9.28	9.30
2	$\ell = 400$	2500	1019	0.73	1.40	1737	14.01	
	$v = 2020$	3000	1106	0.43	1.76	1703	15.69	
		3500	1209	0.56	1.82	1918	5.05	5.05
3	$\ell = 113$	3000	1234	0.54	1.50	1936	16.62	
	$v = 2322$	3500	1375	0.69	1.49	2205	5.04	
		4000	1485	0.56	1.72	2277	1.94	1.94
] 4	$\ell = 99$	5500	912	0.70	2.77	1983	17.41	
	$v = 2401$	6000	1020	0.60	2.90	2069	13.83	
		8000	1139	0.35	3.36	2383	0.75	0.75
5	$\ell = 99$	2500	711	0.63	2.15	1059	26.46	
	$v = 1440$	3000	769	0.41	2.35	1402	2.64	
		3500	837	0.53	2.61	1501	4.24	4.24
6	$\ell = 504$	4500	2061	0.64	1.17	2920	4.39	
	$v = 3054$	5000	2263	0.88	1.18	3335	9.20	
		5500	2452	0.84	1.17	3051	0.10	0.10
7	$\ell = 386$	3500	1825	0.46	1.00	2741	8.08	
	$v = 2982$	4000	1925	0.38	1.17	2833	5.00	
		4500	2159	0.97	1.18	3019	1.24	1.24
8	$\ell = 302$	3500	1473	0.78	1.29	2482	4.21	
	$v = 2591$	4000	1589	0.55	1.36	2396	7.53	
		4500	1704	0.58	1.43	2544	1.81	1.81
9	$\ell = 284$	4000	1582	0.60	1.19	2943	10.06	
	$v = 3272$	4500	1660	0.39	1.85	3740	14.30	
		5000	1720	0.33	1.33	3195	2.35	2.35
10	$\ell = 124$	5000	1840	0.61	1.67	2990	13.11	
	$v = 3441$	5500	1900	0.33	1.71	3208	6.78	
		6000	1997	0.56	1.80	3339	2.96	2.96
						Average deviation (largest M taken)		2.97

Table 7.3 Approximations of v for selected values of M.

quality of approximations: MRPL9: 28.8% deviation vs 14.7% (Table 7.3); PAX7: 18.9% vs 13.8%; GMEB1: 22.9% vs 5.0%.

As mentioned already at the beginning of the section, the random selection of M samples may have an impact on the quality of approximations even

for increasing values of M (although, only marginally; see in Table 7.3 sequences No 5 and No 8). Similarly, there are slight variations in the quality of approximations for fixed values of M, which can be seen already from the distribution of β_j values. For example, for ALDH4A1 and three random selections of $M = 3,000$ initial secondary structures (one related to entry in 7.3, the distributions of $\beta_1, ..., \beta_{10}$ are shown in Table 7.4.

β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	lm(M)	Δ_a%
654	209	74	31	28	17	15	11	11	9	1106	15.7
See Table 7.3											
828	171	54	40	21	25	16	19	12	7	1232	14.3
813	181	63	35	27	17	16	11	15	10	1229	13.8

Table 7.4 β_j data for independent $M = 3,000$ runs for ALDH4A1.

The larger number of observed local minima slightly affects the quality of the approximation, and the random selection for three of the sample sets causes a much larger β_1 value and a smaller β_2, i.e., more local minima are detected by a single initial secondary structure, whereas for the first example more local minima are 'visited' by two initial secondary structures.

We observed that the number of transitions within attractions basins towards local minima is relatively small. For example, for sequence No 1 (MRPL9) and the run from 7.3 with $M = 6,000$, the total number of descent steps is 18,705, which means on average 3.12 transitions, and the percentage of secondary structures visited is 14.74% in relation to $|\mathsf{C}_{\Delta E}|$. For sequence No 4 (PAX7) with $M = 6,000$ the total number is 19,724, which means on average 3.29 steps, and the percentage of visited secondary structures is 23.28%. For sequence No 3 (GMEB1) with $M = 3,500$ the values are 5,143 transitions, 1.47 on average, and 41.54% of secondary structures visited. Of course, as explained in Section 7.2.2, the parameters for GMEB1 were chosen in such a way that the partial landscape has the lowest value of $|\mathsf{C}_{\Delta E}|/\nu$ and about half the ratio of sequence No 2, which explains the relatively high percentage of 41.54%.

For each individual $S \in \mathsf{M}$ of length n, the time for executing a single step of steepest descent can be roughly upper bounded by $O\left(n^2 E_n\right)$, where $O\left(n^2\right)$ covers the number of potential neighbourhood transitions and E_n is the time required for free energy calculations for each individual transition. Since we are using `RNAlocmin` for the steepest descent procedure, where the

energy is updated only for base pair bindings affected by the transition, we can assume $E_n = (1)$. Furthermore, let D denote the maximum length of a steepest descent pathway from elements of M within their respective attraction basin to the corresponding metastable conformation. For example, if all secondary structures from the sample set M are at most 10 kcal/mol above the minimum free energy conformation, and if the free energy of secondary structures is calculated with a precision of 0.1 kcal/mol (standard setting in `RNAsubopt`), then the value of D does not exceed 100. Therefore, the number of steps required for finding all metastable conformations associated with all $S \in \mathsf{M}$ is bounded by $O(M \cdot D \cdot n^2)$, where $M = |\mathsf{M}|$. The result is a list L of M secondary structures representing local minima, with some of the minima potentially being identical.

Given the list L, the set of values $\{\beta_j\}_{j=1}^{M}$ is calculated. If the secondary structures are encoded as natural numbers of length n ("·" $\rightarrow 0$; "(" $\rightarrow 1$; ")" $\rightarrow 2$), one can apply a standard sorting algorithm that returns a sorted list L' where identical secondary structures appear as consecutive elements of the sorted list. The number of comparisons required is $O(M \cdot \log M)$, with a linear complexity $O(n)$ of each comparison. From L' one can then identify the values of individual β_j. The procedure also provides the information about $j_{\max}$, which is the maximum j such that $\beta_j > 0$. Taking into account the exponential upper bound for the number of all secondary structures of length n [57], it is justified to assume $\log M \leq O(n)$, which leads to an upper bound of $O(M \cdot n^2)$ for the number of basic operations required for calculating the set of values $\{\beta_j\}_{j=1}^{j_{\max}}$.

Processing the set of values $\{\beta_j\}_{j=1}^{j_{\max}}$ according to (A) until (E) is determined by $j_{\max}$, the γ-range G, the r-range R, and by their corresponding step-size δ_γ and δ_r. The time complexity is upper bounded by $O(j_{\max} \cdot H/\delta)$ for $H = \max\{G, R\}$; $\delta = \delta_\gamma = \delta_r$.

Altogether, the overall run-time is dominated by $O(M \cdot D \cdot n^2)$. If we assume $M \sim \nu$, i.e., the best approximation is selected for M having the same order of magnitude as ν (which is the case in our computational experiments, see Table 7.3), then the upper bound turns to $O(\nu \cdot D \cdot n^2)$.

In terms of real processing times, the generation of the information about ν is in the region of a few minutes (first `RNAsubopt`, then `barriers`) for the folding landscape data from Table 7.2. The steepest descent by using `RNAlocmin` and the generation of the β_j-values is in the region of a few seconds for the values of M under consideration. The evaluation of the β_j-file

with the return of ν_a terminates after less than 0.5 sec for the data reported in Table 7.3.

7.4 Impact of descent strategy on approximation results

As approximation results are derived from information about the distribition of β_j, i.e. the number of local minima having collected exactly j of M landscape elements, which is calculated after applying a descent procedure over M. In this section, we analyse further the impact of the chosen descent procedure on approximation results and run-time on large energy offsets. Table 7.5 shows for the same set of sequences as given in Table 7.2 large partial landscapes (see also Chapter 6).

No	Gene Name	ℓ	ΔE	$\lvert C_{\Delta E}\rvert$	ν	$\lvert C_{\Delta E}\rvert/\nu$
1	PAX7	99	16.2	14,340,878	50,861	282.0
2	OXT	99	15.0	14,164,430	74,426	190.3
3	GMEB1	113	10.5	15,845,050	466,093	34.0
4	LIG3	124	13.0	15,525,022	317,284	48.9
5	CBR1	284	6.0	10,987,435	643,999	17.1
6	HTR3E	302	9.0	15,095,701	533,316	28.3
7	HLA-G	386	4.2	15,791,146	906,393	17.4
8	ALDH4A1	400	5.4	15,186,200	540,609	28.1
9	MRPL9	407	6.2	14,023,048	41,979	334.0
10	AQP5	504	5.5	11,173,352	714,812	15.6

Table 7.5 Partial energy landscapes for larger ΔE values.

The worst deviation result over all ten sequences is CBR1 where the number of local minima is underestimated by 14.33% (gradient), 15.24% (random) and 9.10% (first lower), see Table 6.3 for . The maximum difference in deviation between the three descents occurs for sequence ALDH4A1 where random deviation = 9.38% and first lower deviation = 3.11%. The best deviations are achieved for the longest sequence AQP5 where gradient deviation = 0.91%, random = 0.04% and first lower = 3.21%. For seven of the ten sequences first lower descent results in the best deviation values. For the remaining three sequences, (HTR3E, HLA-G and AQP5) the maximum difference in deviation compared to gradient is 2.3%.

No	ℓ	ν	M	M/ν	Gradient %-deviation	Random %-deviation	First %-deviation
1	99	50,861	1.0	19.7	8.68	7.74	4.86
2	99	74,426	1.0	13.4	1.09	2.93	0.20
3	113	466,093	2.0	4.3	3.57	2.21	3.26
4	124	317,284	1.0	3.2	12.07	14.72	11.57
5	284	643,999	1.5	2.3	14.33	15.24	9.10
6	302	533,316	1.5	2.8	7.15	6.94	7.33
7	386	906,393	2.5	2.8	2.03	6.57	3.77
8	400	540,609	1.5	2.8	3.52	12.49	3.11
9	407	41,979	1.0	23.8	7.16	3.73	3.40
10	504	714,812	2.0	2.8	0.91	0.04	3.21
				Average	6.05	7.26	4.98
				Standard Deviation	4.64	5.34	3.37

Table 7.6 Approximation results: Percentage deviations to the number of local minima ν reported by `Barriers` for M (million) randomly selected structures.

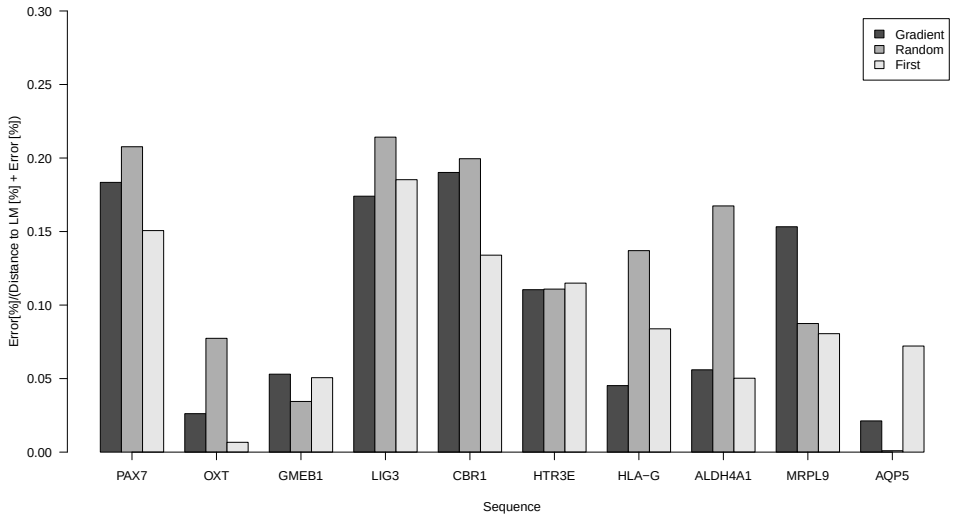

Fig. 7.3 Approximation Error: quality of approximation for each descent procedure.

For the three longest sequences with M ranging from 1 to 2 million, the `RNAsubopt` run-time exceeds the run-time of descent[1]. For example, the run-time of `RNAsubopt` on MRPL9 is approximately 15.12 minutes, whereas

[1] All runtimes generated using Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz and 32GB RAM.

descent for $M = 1$ million conformations is about 5.18 minutes (gradient), 4.15 (random) and 3.95 (first lower), see Table 7.7 and Figure 7.4. As descent is the dominating run-time factor for the approximation approach the run-time for longer sequences could possibly be improved by replacing `RNAsubopt` with a random secondary structure generator.

No	ℓ	$M \times 10^6$	`RNAsubopt` + `Barriers` time	Gradient time	Random time	First time
1	99	1.0	7.71	8.48	6.75	5.38
2	99	1.0	7.18	5.44	5.45	5.41
3	113	2.0	15.11	5.74	8.45	5.09
4	124	1.0	12.22	5.65	5.11	4.94
5	284	1.5	24.39	13.29	12.50	12.14
6	302	1.5	40.48	14.78	12.84	12.82
7	386	2.5	69.37	22.14	21.80	20.06
8	400	1.5	63.55	17.08	16.36	15.64
9	407	1.0	55.33	20.71	19.83	20.12
10	504	2.0	63.38	24.99	24.24	23.80
		Total	358.72	138.3	133.33	125.4

Table 7.7 Total run-time time in minutes of `RNAsubopt` + descent method + approximation heuristic.

If approximate solutions are sufficient, as in the comparison of `RNAlocopt` from [191] with `RNAlocmin` regarding the coverage of local minima within a given time frame (as presented in [194]), then our method provides a run-time advantage over `RNAsubopt` + `Barriers`. Figure 7.4 shows the total run-time required to approximate the number of local minima compared to `RNAsubopt` + `Barriers`. In the approximation approach, the run-time includes running `RNAsubopt` to generate the secondary structures, descent method to calculate the local minima, approximation heuristic. The figure clearly shows that the approximation approach for all three descent methods is an improvement over exhaustively generating and merging of structures by `Barriers`. For example, the total run-time for sequence MRPL: `RNAsubopt` = 15.12 minutes + random descent for $M = 1$ million = 4.15 + approximation heuristic = 0.56 giving a total time of 19.83 minutes. The run-time of `RNAsubopt`+ `Barriers` for MRPL9 = 53.33 minutes resulting in an improvement of 33.5 minutes with a deviation of 3.73% for random descent.

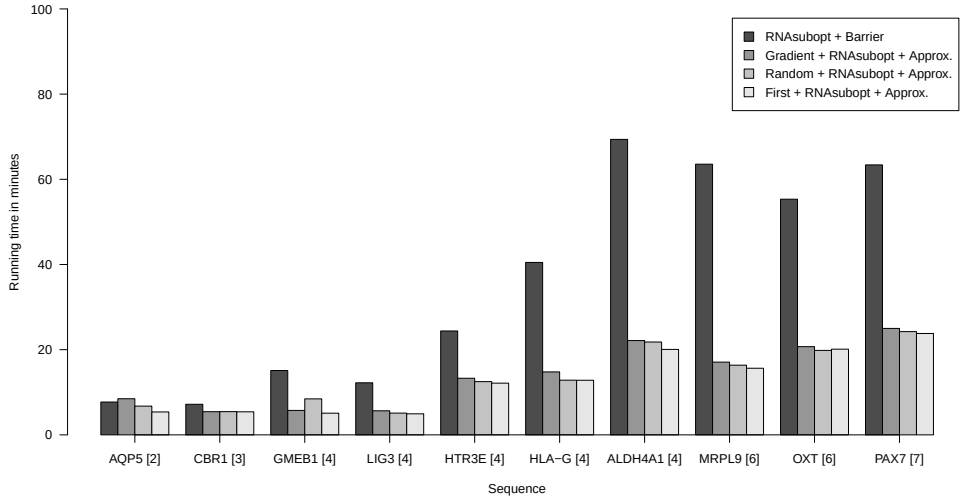

Fig. 7.4 Approximation time: Total running time in minutes of `RNAsubopt` + descent method + approximation heuristic compared to `RNAsubopt` + `Barriers`.

7.5 Conclusions

In this chapter we presented a new sampling method to approximate the number of local minima in partial energy landscapes. On the ten sequences we analysed, the new pooling procedure along with the separation of γ_a and r_a calculations resulted in an average deviation of about 3% from the true number of local minima. The approximation procedure is applicable in a scenario where no *a priori* knowledge about the number of local minima is available, as discussed, for example, in Kucharík *et al.* [194] where the authors discuss the dynamics of RNA folding over energy landscapes. Furthermore, we evaluated impact of three descent methods on approximation quality and run-time over large partial landscapes. We obtained on average a total run-time improvement of 3.7 hours compared to exhaustive generation and filtering of structures with an average deviation of 6.05% (gradient), 7.26% (random) and 4.98% (first lower) over all ten sequences.

Chapter 8

Conclusions

8.1 Summary

A fundamental principle of structural biology is that sequence encodes structure and in turn structure provides insights into function. The rate at which RNA structures are being determined experimentally lags significantly behind that of proteins. The ultimate goal of RNA and protein structure prediction is to determine their three dimensional structures. However, determining RNA's three dimensional structure is currently too computationally demanding. Computational secondary structure predictions and analyses are most commonly based on thermodynamic stability where the focus is on the single minimum free energy conformation. However, it is now commonly acknowledged that *in vivo* RNAs may not always fold into their minimum free energy conformations and may instead fold into an ensemble of structural states. Consequently, this suggests that the information flow for RNAs is better described by sequence $\rightarrow$ *folding landscape* $\rightarrow$ structure $\rightarrow$ function.

In this thesis we have highlighted the importance and need for computational methods and analyses that take into consideration metastable RNA structure. In the first contribution chapter (Chapter 4), we analysed how Single Nucleotide Polymorphisms (SNPs) can affect the RNA secondary structure ensemble. And, if occurring within a microRNA binding site, can result in a change in acessibility. We identified from published literature strong expression level analyses investigating [mRNA; SNP; miRNA] associations in the context of disease risk. By analyses of metastable conformations, we identified three parameters from the RNA folding landscape that provide supporting information for the experimentally observed differences in expression of alleles

defined by a SNP. Analyses of MFE only structures as commonly done within mRNA-microRNA studies does not provide sufficient information on binding site accessibility.

In the second contribution chapter (Chapter 5), we applied our findings into a microRNA target site prediction tool `RNAStrucTar` that takes into consideration metastable binding site accessibility. And, analysed the tool on set of 20 [mRNA; SNP; miRNA] instances identified from published literature. We found the prediction tool correctly classifies the allele where microRNA is reported to have stronger binding for 16 of the instances. In comparison to similar prediction tool, `PITA` correctly identifies 13 and `STarMir` 14 instances, suggesting that the inclusion of metastable binding site structure provides useful information for microRNA target site predictions.

In the third contribution chapter (Chapter 6), we firstly discussed the importance of descent methods in RNA energy landscape. We then compared deterministic and random descent methods over partial RNA folding landscapes. In our comparison, we focused on the coverage of metastable structures and differences in run-times. Moreover, for the two nongradient methods we analysed for partial energy landscapes induced by ten different RNA sequences, we obtained that the number of observed local minima is on average larger by 7.3% and 3.5%, respectively. The run-time improvement is approximately 16.6% and 6.8% on average over the ten partial energy landscapes.

In the fourth contribution chapter (Chapter 7), a new heuristic method was proposed based on the general framework devised by Garnier and Kallel for approximating the number of local minima in partial RNA folding landscapes. Over ten RNA sequences, our heuristic method achieves for best approximations on average a deviation below 3.0% from the true number of local minima. We then analysed the impact of descent strategy on the approximation heuristic over ten large partial energy landscapes. The approximation heuristic achieves an average deviation of 6% when using steepest or gradient descent, 5% first-lower descent and 7.26% when using random descent. And, a total run-time in improvement of 3.7 hours over all ten sequences in comparison to exhaustive generation and filtering by `RNAsubopt` + `Barriers`.

8.2 Future outlook

The combination of decreasing genomic sequencing costs and growing interest within industry on potential applications of RNA as a new treatment for disease offers much hope to revolutionise medicine. Of special clinical interest are microRNAs which are now strongly associated with disease development and progression, especially cancer pathways where it has been shown that dysregulation of a single microRNA is sufficient to cause cancer. However, decoding the functional complexity of RNA is a huge task with many unanswered questions. For example, in microRNA biogenesis it is unknown how a pre-miRNA helix separates into single strands and binds to the RNA induced silencing complex. An ideal RNA structure prediction tool would take into consideration the following:

- **Three dimensional interactions**
  Currently there are insufficient parameters to accurately model RNAs tertiary interactions and the lack of experimentally verified RNA structures make it difficult to determine useful three dimensional motifs to make full predictions.

- **Co-transcriptional folding**
  Co-transcriptional folding is generally acknowledged as how RNA folds. Computationally, a key problem is reducing the conformational space and a better biological understanding of co-transcriptional folding could be used to reduce such space by means of folding pathways [203]. Geary *et al.* use a co-transcriptional folding framework to constrain the folding pathway of RNA origami structures [204, 205].

- **The dynamics of folding**
  Related to co-transcriptional folding is the kinetics of folding and refolding. If RNA does not fold to a single static conformation then the kinetics or dynamics of RNA folding need to be taken into consideration. The kinetics of folding is particularly important for riboswitches and RNA thermometers.

- **Cells environment**
  RNAs typically do not fold in isolation, instead interact with many other molecules such as metal ions which can help stabilise structure.

Each of the above present many computationally demanding problems. Computational prediction of structure relies upon experimental observations that can be incorporated into new tools. However, thus far experimental RNA structure methods lag significantly behind those of proteins and therefore there is a lack of strong experimental data to derive new prediction methods and to verify existing proposed methods. However, as new and current experimental techniques targeting RNA structure advance, such as [122, 123], it is likely they will provide new insights into the fundamental rules governing RNA structure formation and their interaction with other molecules allowing for more focused prediction methods.

One area of potential future research, would be to modify `MSbind` to analyse the proximity and accessibility of microRNA and RNA-binding protein (RBP) sites. Over recent years, there has been a growing number of publications detailing cooperation and competition between microRNAs and RNA-binding proteins. For example, it is well documented that let-7 repression of c-myc requires the RBP HuR. And, miRNA-125b repression of tumour suppressor P53 can be blocked by HuR likely due to overlapping binding sites. See Appendix D for list of other interesting cases and references. Using resources such as:

- `AREsite2` [206], a database of AU-rich 3' UTRs.
- `CLIPdb` [207], a database of experimentally determined RBP sites.
- `miRTarBase` [185], a databases of experimentally validated microRNA-target interactions.

In the case of investigating RBP-miRNA cooperation, examining AU-rich sequence sites do not provide insight into how close in a folded state the RBP site is to the microRNA. That is, the number of shared nucleotide pairings if for example the two binding sites occur at opposite sides of a helix. It would be of interest to map validated microRNA sites and RBPs sites over deep metastable structures to determine: (1) any overlapping RBP-miRNA sites, (2) the accessibility of the binding sites, and (3) the structural proximity of all microRNA sites to all RBP sites.

References

[1] National Human Genome Research Institute. DNA sequencing costs from the NHGRI Genome Sequencing Program. https://www.genome.gov/sequencingcosts/, accessed: March 2016.

[2] National Human Genome Research Institute. The human genome project completion. https://www.genome.gov/11006943/, accessed: March 2016.

[3] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447:799–816, 2007.

[4] The Wellcome Trust Sanger Institute. GENCODE project. http://www.gencodegenes.org/stats.html, accessed: March 2016.

[5] Kirkwood Land and Lisa A. Wrischnik. Basic biology of Trichomonas Vaginalis: current explorations and future directions. *Sex Transm Infect*, 89(6):416–417, 2013.

[6] The ENCODE Project Consortium. Encode explorer. http://www.nature.com/encode/, accessed: March 2016.

[7] Palazzo AF and Lee ES. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*, 6(2), 2015.

[8] Morris KV and Mattick JS. The rise of regulatory RNA. *Nature Reviews Genetics*, 15:423–437, 2014.

[9] Day L, Abdelhadi Ep Souki O, Albrecht A. A, Steinhöfel, K. Accessibility of microRNA binding sites in metastable RNA secondary structures in the presence of SNPs. *Bioinformatics*, 30(3):343–353, 2014.

[10] Abdelhadi Ep Souki O, Day L, Albrecht AA and Steinhöfel K. MicroRNA Target Prediction Based Upon Metastable RNA Secondary Structures. *Bioinformatics and Biomedical Engineering*, 9044:456–467, 2015.

[11] Day L, Abdelhadi Ep Souki O, Albrecht AA and Steinhöfel K. Random versus deterministic descent in RNA energy landscape analysis. *Advances in Bioinformatics*, (9654921), 2016.

[12] Albrecht AA, Day L, Abdelhadi Ep Souki O, Steinhöfel, K. A new heuristic method for approximating the number of local minima in partial RNA energy landscapes. *Computational Biology and Chemistry*, 60:43–52, 2016.

[13] Mizuno H, Sundaralingam M. Stacking of Crick wobble pair and Watson-Crick pair: stability rules of G-U pairs at ends of helical stems in tRNAs and the relation to codon-anticodon wobble interaction. *Nucleic Acids Res*, 5:4451–4461, 1978.

[14] Masquida B and Westhof E. On the wobble GU and related pairs. *RNA*, 6(1):9–15, 1978.

[15] Miller S, Jones LE, Giovannitti K, Piper D, Serra MJ. Thermodynamic analysis of 5 and 3 single- and 3' double-nucleotide overhangs neighboring wobble terminal base pairs. *Nucleic Acids Research*, 36(17):5652–5659, 2008.

[16] Crick F. On Protein Synthesis. *The Symposia of the Society for Experimental Biology*, 12(13):138–162, 1958.

[17] Crick F. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[18] Cech TR, Zaug AJ and Grabowski PJ. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27(3):487–496, 1981.

[19] Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983.

[20] Cech TR. A model for the RNA-catalyzed replication of RNA. *Proc Natl Acad Sci*, 83(12):4360–4363, 1986.

[21] Higgs PG and Lehman N. The RNA World: molecular cooperation at the origins of life. *Nature Reviews Genetics*, 16:7–17, 2015.

[22] Waterman, M. S. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies*, 1:167–212, 1978.

[23] Tuschl T, Gohlke C, Jovin T. M, Westhof E, and Eckstein F. A three dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science*, 266(5186):785–789, 1994.

[24] LLC Schrödinger. The pymol molecular graphics system. http://www.pymol.org/, accessed: March 2016.

[25] José AC., Westhof E. The Dynamic Landscapes of RNA Architecture. *Cell*, 136(4):604–609, 2009.

[26] Gruber A., Bernhart S., Hofacker I., Washietl S. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, 9(122), 2008.

[27] Westhof E., Masquida B., Jossinet F. Predicting and Modeling RNA Architecture. *Cold Spring Harb Perspect Biol*, 3(2), 2011.

[28] Tinoco I Jr and Bustamante C. How RNA folds. *J. Mol. Biol.*, 293(2):271–281, 1999.

[29] Ha M, Kim N. Regulation of microRNA biogenesis. *Nature Rev Mol Cell Bio*, 15, 2014.

[30] Lee RC, Feinbaum RL and Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

[31] Rougvie AE. Control of developmental timing in animals. *Nature Reviews Genetics*, 2(9):690–701, 2001.

[32] Ambros V. microRNAs: tiny regulators with great potential. *Cell*, 107(7):823–826, 2001.

[33] Kozomara A, and Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2014.

[34] Hesketh R. *Introduction to cancer biology: a concise journey from epidemiology through cell and molecular biology to treatment and prospects*. Cambridge University Press, 2013.

[35] Calin GA, Dumitry CD, Shimizu M. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci*, 99(24):15524–15529, 2002.

[36] Johnson SM, Grosshans H, Shingara J, Byrom M, Jarvis R, Cheng A, Labourier E, Reinert KL, Brown D, Slack FJ. RAS is regulated by the let-7 microRNA family. *Cell*, 120(5):635–647, 2005.

[37] Kwan JYY, Psarianos P, Bruce JP, Yip KW, Liu F. The complexity of microRNAs in human cancer. *J Radiat Res*, doi: 10.1093/jrr/rrw009(EPub ahead of print), 2016.

[38] Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*, 1(15004), 2016.

[39] Lu et al. MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838, 2005.

[40] Medina PP, Nolde M, Slack FJ. OncomiR addiction in an *in vivo* model of microRNA-21-induced pre-B-cell lymphoma. *Nature*, 467:89–90, 2010.

[41] Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res*, 42(Database Issue D-1070-4), 2014.

[42] Cho S, Jun Y, Lee S, Choi HS, Jung S, Jang Y, Park C, Kim S, Lee S, Kim W. miRGator v2.0: an integrated system for functional investigation of microRNAs. *Nucleic Acids Res*, 39(Database Issue D-158-62), 2011.

[43] Ling H, Fabbri M, Calin GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nature Reviews Drug Discovery*, 12:847–865, 2013.

[44] Rani V, Yadav T, Mishra S. Exploring miRNA based approaches in cancer diagnostics and therapeutics. *Critical Rev in Oncology/Hematology*, 98:12–23, 2016.

[45] Janssen HLA et al. Treatment of HCV Infection by Targeting MicroRNA. *N Engl J Med*, 368(18):1685–1694, 2013.

[46] Ottosen UA, et al. In vitro antiviral activity and preclinical and clinical resistance profile of miravirsen, a novel anti-hepatitis C virus therapeutic targeting the human factor miR-122. *Antimicrob. Agents Chemother*, 59:599–608, 2015.

[47] Bouchie A. First microRNA mimic enters clinic. *Nature Biotechnology*, 31(577), 2013.

[48] Wong E and Goldberg T. Mipomersen (Kynamro): A Novel Antisense Oligonucleotide Inhibitor for the Management of Homozygous Familial Hypercholesterolemia. *Pharmacy and Therapeutics*, 39(2):119–122, 2014.

[49] Clancy JL, Nousch M, Humphreys DT, Westman BJ, Beilharz TH, Preiss T. Methods to analyze microRNA-mediated control of mRNA translation. *Methods Enzymol*, 431:83–111, 2007.

[50] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141:129–141, 2010.

[51] Licatalosi D, Mele A, Fak J, Ule J, Kayikci M, Chi S, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456:464–469, 2008.

[52] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153:654–665, 2013.

[53] Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, 2003.

[54] Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math*, 45:810–825, 1985.

[55] Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):90–98, 2006.

[56] Bore P., Dengle B., Tinoco I., Uhlenbeck O. Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology*, 86:843–853, 1974.

[57] Waterman, M. S, Smith T. F. RNA secondary structure: A complete mathematical analysis. *Math Biosci.*, 42:257–266, 1978.

[58] Zuker M, Sankoff D. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–624, 1984.

[59] Nussinov R, Pieczenik G, Griggs J, Kleitman D. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1), 1978.

[60] Nussinov R, Jacobson A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci*, 77(11):6309–6313, 1980.

[61] Tinoco I, Uhlenbeck O. C., Levine M. D. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, 1971.

[62] Freier S. M , Kierzek R, Jaeger J. A, Sugimoto A, Caruthers M. H, Neilson T, Turner D. H. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences.*, 83(24), 1986.

[63] Jaeger J, Turner D H, Zuker M. Improved predictions of secondary structures for RNA. *Proceedings of the National Academy of Sciences*, 86, 1989.

[64] Andronescu M, Condon A, Holger HH, Mathews DH, Murphy KP. Computational approaches for RNA energy parameter estimation. *RNA*, 16(12):2304–2318, 2010.

[65] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1), 2011.

[66] Turner D. H, Mathews D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucl. Acids Res.*, 38(suppl 1):D280–D282, 2010.

[67] Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucl. Acids Res.*, 34(17):4912–24, 2006.

[68] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

[69] Hofacker I. L, Fontana W, Stadler P. F, Bonhoeffer L. S, Tacker M, Schuster P. Folding and Comparison of RNA Secondary Structures. *Chemical Monthly*, 125(2):440–445, 1994.

[70] Lyngso R. B, Zuker M, Pedersen C. N. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15(6):440–445, 1999.

[71] Markham NR, Zuker M. UNAFold, software for nucleic acid folding and hybridization. *Methods Mol Biol*, 453, 2008.

[72] Reuter JS, Mathews DH. RNAstructure: software for secondary structure prediction and analysis. *BMC Bioinformatics*, 11(129), 2010.

[73] Eddy SR. How do RNA folding algorithms work? *Nat. Biotech.*, 22:1457–1458, 2004.

[74] Lyngso RB, Pedersen CN. RNA pseudoknot prediction in energy-based models. *J Comput Biol*, 7(3–4):409–427, 2000.

[75] Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, 285(5):2053–2068, 1999.

[76] Dirks RM, Pierce NA. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem*, 24(13):1664–1677, 2003.

[77] Pan T, Sosnick T. RNA folding during transcription. *Annual review of biophysics and biomolecular structure*, 35, 2006.

[78] Al-Hashimi HM, Walter NG. RNA dynamics: It is about time. *Curr Opin Struct Biol*, 18:321–29, 2008.

[79] Lai D, Proctor JR, Meyer IM. On the importance of cotranscriptional RNA structure formation. *RNA*, 19(11):1461–73, 2013.

[80] Heilman-Miller SL, Woodson SA. Effect of transcription on folding of the Tetrahymena ribozyme. *RNA*, 9(6):722–33, 2003.

[81] Levinthal C. How to Fold Graciously. *Mössbaun Spectroscopy in Biological Systems Proceedings*, 67(41):22–24, 1969.

[82] Garst AD, Edwards AL, Batey RT. Riboswitches: Structures and Mechanisms. *Cold Sping Harb Perspect Biol*, 3(6), 2011.

[83] Reining A, Nozinovic S, Schlepckow K, Buhr F, Fürtig B, Schwalbe H. Three-state mechanism couples temperature sensing in riboswitches. *Nature*, 499(7458):355–359, 2013.

[84] Flores R, Serra P, Minoia S, Navarro B. Viroids: From Genotype to Phenotype Just Relying on RNA Sequence and Structural Motifs. *Front Microbiol*, 3(217), 2012.

[85] Solomatin SV, Greenfeld M, Chu S, Herschlag D. Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature*, 463(4281):681–684, 2010.

[86] Chursov A, Kopetzky SJ, Bocharov G, Frishman D, Shneider A. RNAtips: analysis of temperature-induced changes of RNA secondary structure. *Nucl. Acids Res*, 41(Web Issue):W486–W491, 2013.

[87] Righetti F, Narberhaus F. How to find RNA thermometers. *Front Cell Infect Microbiol*, 4(132), 2014.

[88] Mustoe AM, Brooks CL, Al-Hashimi HM. Hierarchy of RNA functional dynamics. *Annual reviews of biochemistry*, 83:441–466, 2014.

[89] Zuker M. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.

[90] Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structure. *Biopolymers*, 49(2):145–165, 1999.

[91] Waterman MS, Byers T. A dynamic programming algorithm to find all solutions in a neighbourhood of the optimum. *Math Biosci.*, 77:179–188, 1985.

[92] Stone JW, Bleckley S, Lavelle S, Schroeder SJ. A Parallel Implementation of the Wuchty Algorithm with Additional Experimental Filters to More Thoroughly Explore RNA Conformational Space. *PLoS ONE*, 10(2), 2015.

[93] Flamm C, Fontana W, Hofacker IL, Schuster P. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, 2000.

[94] Chen SJ, Dill KA. RNA folding energy landscapes. *PNAS*, 97(2):646–651, 2000.

[95] Morgan S, Higgs P. Barrier heights between ground states in a model of RNA secondary structure. *J. Phys. Math. and Gen.)*, 31(14):3153–3170, 1998.

[96] Manuch J, Thachuk C, Stacho L, Condon A. NP-completeness of the direct energy barrier problem without pseudoknots. *DNA Computing and Molecular Programming, LNCS*, 5877:106–115, 2009.

[97] Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M. Design of multistable RNA molecules. *RNA)*, 7:254–265, 2001.

[98] Flamm C, Hofacker IL, Stadler PF, Wolfinger MT. Barrier Trees of Degenerate Landscapes. *Z. Phys. Chem*, 216:155–173, 2002.

[99] Dotu I, Lorenz WA, Hentenryck PV, Clote P. Computing folding pathways between RNA secondary structures. *Nucleic Acids Res.)*, 38(5):1711–1722, 2010.

[100] Zucker M. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 26(6–7):1105–1119, 1990.

[101] Cupal J, Hofacker IL, Stadler PF. Dynamic programming algorithm for the density of states of RNA secondary structures. *Computer Science and Biology 96 (German Conference on Bioinformatics)*, pages 1157–1166, 1996.

[102] Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. *Nucl. Acids Res*, 31(24):7280–7301, 2003.

[103] Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

[104] Tafer H, Hofacker IL. RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics*, 24(22):2657–2663, 2008.

[105] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. *Genome Biol*, 5(1), 2003.

[106] Bartel DP. MicroRNA: target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[107] Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19:92 – 105, 2009.

[108] Aneres SL, Martinez J, Schroeder R. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, 130(1):101–112, 2007.

[109] Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455:58–63, 2015.

[110] Zhang H, Artiles KL, Fire AZ. Functional relevance of "seed" and "non-seed" sequences in microRNA-mediated promotion of C. elegans developmental progression. *RNA*, 21:1980–1992, 2015.

[111] Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4(e05005), 2015.

[112] Muckstein U, Tafer H, Hackermuller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10), 2006.

[113] Bernhart SH, Tafer H, Muckstein U, Flamm C, Stadler PF, Hofacker IL. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology*, 1(1), 2006.

[114] Andronescu M, Zhang ZC, Condon A. Secondary Structure Prediction of Interacting RNA Molecules. *Journal of Molecular Biology*, 345(5):987–1001, 2005.

[115] Bernhart SH, Muckstein U, Hofacker IL. RNA Accessibility in cubic time. *Algorithms Mol Biol*, 6(3), 2011.

[116] Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. Potent effect of target structure on microRNA function. *Nat. Struc. Mol. Biol.*, 14(4):287–294, 2007.

[117] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39, 2007.

[118] Alkan C, Karakoc E, Nadeau JH, Sahinalp SC, Zhang K. RNA-RNA interaction prediction and antisense RNA target search. *Journal of Computational Biology*, 13(2):267–282, 2006.

[119] Lai D, Meyer IM. A comprehensive comparison of general RNA–RNA interaction prediction methods. *Nucleic Acids Res*, 44(7), 2016.

[120] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, 2003.

[121] Cloonan N. Re-thinking miRNA-mRNA interactions: Intertwining issues confound target discovery. *BioEssays*, 37(4), 2015.

[122] Lorenz R, Wolfinger MT, Tanzer A, Hofacker IL. Predicting RNA secondary structures from sequence and probing data. *Methods*, In press(Epub), 2016.

[123] Lorenz R, Luntzer D, Hofacker IL, Stadler PF, Wolfinger MT. SHAPE directed RNA folding. *Bioinformatics*, 32, 2016.

[124] 1000 Genomes Project Consortium. An Integrated Map of Genetic Variation from 1,092 Human Genome. *Nature*, 491:55–65, 2012.

[125] Salzman DW, Weidhaas JB. SNPing cancer in the bud: microRNA and microRNA-target site polymorphisms as diagnostic and prognostic biomarkers in cancer. *Pharmacol Ther*, 137:55–63, 2013.

[126] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29:308–311, 2001.

[127] Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Gent*, 17(9):502–510, 2001.

[128] Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, 8(12), 2012.

[129] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[130] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–D1006, 2014.

[131] Allison AC. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *British Medical Journal*, 1(4857):290–294, 1954.

[132] Rockett et al. Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicentre study. *Nat Genet*, 46(11):1197–1204, 2014.

[133] Goncalves BP, Gupta S, Penman BS. Sickle haemoglobin, haemoglobin C and malaria mortality feedbacks. *Malar J*, 15(26), 2016.

[134] Praveen S, Collins FS. MicroRNA target site polymorphisms and human disease. *Trends in genetics*, 24(10):489–497, 2008.

[135] Chin et al. A SNP in a let-7 microRNA complementary site in the KRAS 3UTR Increases Non-Small Cell Lung Cancer Risk. *Cancer Res.*, 68(20):8535–8540, 2008.

[136] Margalit H, Shapiro BA, Oppenhein AB, Maizel Jr JV. Detection of common motifs in RNA secondary structures. *Nucleic Acids Res.*, 17(12):4829–4845, 1989.

[137] Churkin A, Barash D. RNAmute: RNA secondary structure mutation analysis tool. *BMC Bioinformatics*, 7(221), 2006.

[138] Waldispühl J, Devandas S, Berger B, Clote P. Efficient Algorithms for Probing the RNA Mutation Landscape. *PLoS Comput Biol*, 4(8), 2008.

[139] Halvorsen M, Martin JS, Boradaway Sm Laederach A. Disease-Associated Mutations That Alter the RNA Structural Ensemble. *PLoS Genet.*, 6(8), 2010.

[140] Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum Mutat.*, 34(4), 2013.

[141] Nicoloso et al. Single Nucleotide Polymorphisms inside microRNA Target Sites Influence Tumor Susceptibility. *Cancer Res.*, 70(7):2789–2798, 2010.

[142] Mallick B, Ghosh Z. A complex crosstalk between polymorphic microRNA target sites and AD prognosis. *RNA Biol.*, 8(4):665–673, 2011.

[143] Haas U, Sczakiel G, Laufer SD. MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3'-UTR via altered RNA structure. *RNA Biol.*, 9(6):924–937, 2012.

[144] Johnson AD. et al. RNA structures affected by single nucleotide polymorphisms in transcribed regions of the human genome. *Web Med. Cent. Bioinf*, 2:WMC001600, 2011.

[145] Martin JS. et al. Structural effects of linkage disequilibrium on the transcriptome. *RNA*, 18:77–87, 2012.

[146] Subkhankulova T, Gilchrist TS, Livesey F. Modelling and measuring single cell RNA expression levels find considerable transcriptional differences among phenotypically identical cells. *BMC Genomics*, 9(268), 2008.

[147] Arvey A, et al. Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.*, 6(363), 2010.

[148] Salmena, L. et al. A ceRNA hypothesis: The Rosetta Stone of a hidden RNA language? *Cell*, 146:353–358, 2011.

[149] Mullokandov G, et al. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat. Methods*, 9(8):840–846, 2012.

[150] Garcia DM, et al. Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lsy-6* and other microRNAs. *Nat. Struct. Mol. Biol.*, 18(10):1139–46, 2011.

[151] Larsson E, Sander C, Marks D. mRNA turnover rate limits siRNA and microRNA efficacy. *Mol. Syst. Biol.*, 6(433), 2010.

[152] Cuccato G, et al. Modeling RNA interference in mammalian cells. *BMC Syst. Biol.*, 9(19), 2011.

[153] Osella M, et al. The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS Comput. Biol.*, 7(3):e1001101, 2011.

[154] Loinger A, et al. Competition between small RNAs: A quantitative view. *Biophys. J.*, 102(8):1712–1721, 2012.

[155] Baker C, et al. Toward a combinatorial nature of microRNA regulation in human cells. *Nucleic Acid Res.*, 40(19):9404–16, 2012.

[156] Ragan C, Zucker M, Ragan M. Quantitative prediction of miRNA–mRNA interaction based on equilibrium concentrations. *PLoS Comput. Biol.*, 7(e1001090), 2011.

[157] Marin R.M., Vaniŏek J. Optimal use of conservation and accessibility filters in microRNA target prediction. *PLoS One*, 7(e32208), 2012.

[158] Muniategui G, et al. Joint analysis of miRNA and mRNA expression data. *Brief. Bioinf.*, doi: 10.1093/bib/bbs028, 2012.

[159] Shirdel E, et al. NAViGaTing the micronome - using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs. *PLoS One*, 6(2), 2011.

[160] Gennarino V, et al. Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Res.*, 22(6), 2012.

[161] Balaga O, et al. Toward a combinatorial nature of microRNA regulation in human cells. *Nucleic Acid Res.*, 40(19):9404–16, 2012.

[162] Johnson E, Srivastava R. Volatility in mRNA secondary structure as a design principle for antisense. *Nucleic Acids Res.*, 41(e43), 2013.

[163] Ding Y, Chan CY, Lawrence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11(8):1157–1166, 2005.

[164] Teo M, et al. The role of microRNA-binding site polymorphisms in DNA repair genes as risk factors for bladder cancer and breast cancer and their impact on radiotherapy outcomes. *Carcinogenesis*, 33(3):581–586, 2012.

[165] Kalabus JL, et al. MicroRNAs differentially regulate carbonyl reductase 1 (CBR1) gene expression dependent on the allele status of the common polymorphic variant rs9024. *Carcinogenesis*, 7(11), 2012.

[166] Kapeller J, et al. First evidence for an association of a functional variant in the microRNA-510 target site of the serotonin receptor type 3E gene with diarrhea predominant irritable bowel syndrome. *Hum. Mol. Genet.*, 17(19):2967–77, 2008.

[167] Tan Z, et al. Allele-specific targeting of microRNAs to HLA-G and risk of asthma. *Amer. J. Hum. Genetics*, 81(4):829–384, 2007.

[168] Kovacs-Nagy R, et al. Association of aggression with a novel microRNA binding site polymorphism in the wolframin gene. *Am. J. Med. Genet. B*, 126B(4):404–12, 2013.

[169] Zwiers A, et al. A variant of the IL-23R gene associated with inflammatory bowel disease induces loss of microRNA regulation and enhanced protein production. *J. Immunology*, 188(4):1573–7, 2012.

[170] Wang L, et al. A miRNA binding site single-nucleotide polymorphism in the 3'-UTR region of the IL23R gene is associated with breast cancer. *PLoS One*, 7(12), 2012.

[171] Zhang L, et al. Functional SNP in the microRNA-367 binding site in the 3'UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification. *Proc. Nat. Acad. Sci. USA*, 108(33):13653–8, 2011.

[172] Wang G, et al. Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. *Genet.*, 82(2):283–9, 2008.

[173] Luo J, et al. A microRNA-7 binding site polymorphism in HOXB5 leads to differential gene expression in bladder cancer. *PLos One*, 7(6), 2012.

[174] Chang W, et al. ORAI1 genetic polymorphisms associated with the susceptibility of atopic dermatitis in Japanese and Taiwanese populations. *PLos One*, 7(1), 2012.

[175] Wang K, et al. MiR-196a binding-site SNP regulates RAP1A expression contributing to esophageal squamous cell carcinoma risk and metastasis. *Carcinogenesis*, 33(11), 2012.

[176] Ye W, et al. The effect of central loops in miRNA:MRE duplexes on the efficiency of miRNA-mediated gene regulation. *PLoS One*, 3(3), 2008.

[177] Robins H, Li Y, Padgett RW. Incorporating structure to predict microRNA. *PNAS USA*, 102(11):4006–4009, 2005.

[178] Cheng M, et al. A microRNA-135a/b binding polymorphism in CD133 confers decreased risk and favorable prognosis of lung cancer in Chinese by reducing CD133 expression. *Carcinogenesis*, 34(10):2292–2299, 2013.

[179] Delay C, Calon F, Mathews P, Hebert S. Alzheimer-speci c variants in the 3' UTR of amyloid precursor protein affect microRNA function. *Mol. Neurodegen.*, 6(1), 2011.

[180] Hikami K, et al. Association of a functional polymorphism in the 3'-untranslated region of SPI1 with systemic lupus erythematosus. *Arthritis and Rheumatism*, 63(3):755–763, 2011.

[181] Li Y, et al. G-A variant in miR-200c binding site of EFNA1 alters susceptibility to gastric cancer. *Mol. Carcinogenesis*, 53(3):219–229, 2014.

[182] Minguzzi S, et al. An NTD associated polymorphism in the 3' UTR of MTHFD1L can aect disease risk by altering miRNA binding. *Human Mutation*, 35(1):96–104, 2014.

[183] Zhang S, et al. REV3L 3' UTR 460 T > C polymorphism in microRNA target sites contributes to lung cancer susceptibility. *Human Mutation*, 32:242–250, 2013.

[184] Bruno AE, Li L, Kalabus JL, Yuzhuo P, Aiming Y, Zihua H. miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3'UTRs of human genes. *BMC Genomics*, 13(1), 2012.

[185] Hsu SD, et al. MiRTarbase update 2014: miRTarBase an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, 42(D1):D78–D85, 2014.

[186] Neupane K, et al. Direct observation of transition paths during the folding of proteins and nucleic acids. *Science*, 352(6282):239–242, 2016.

[187] Flamm C, Hofacker IL. Beyond energy minimization: approaches to the kinetic folding of RNA. *I. Monatsh Chem*, 139(447):447–457, 2008.

[188] Hofacker IL, Flamm C, Heine C, Wolfinger MT, Scheuermann G, Stadler PF. BarMap: RNA folding on dynamic energy landscapes. *RNA*, 16(17):1308–1316, 2010.

[189] Pörschke DP. Model calculations on the kinetics of oligonucleotide double helix coil transitions. Evidence for a fast chain sliding reaction. *Biophysical Chemistry*, 2(2):83–96, 1974.

[190] Mahen EM, Watson PY, Cottrell JW, and Fedor MJ. mRNA secondary structures fold sequentially but exchange rapidly in vivo. *PLoS Biology*, 8(2), 2010.

[191] Lorenz WA, and Clote P. Computing the Partition Function for Kinetically Trapped RNA Secondary Structures. *PLoS ONE*, 6(1), 2011.

[192] Li Y, and Zhang S. Finding stable local optimal RNA secondary structures. *Bioinformatics*, 27(21):2994–3001, 2011.

[193] Huang J, Backofen R, Voß B. Abstract folding space analysis based on helices. *RNA*, 18(12):2135–2147, 2012.

[194] Kucharík M, Hofacker IL, Stadler PF, Qin J. Basin hopping graph: a computational framework to characterize RNA folding landscapes. *Bioinformatics*, 30(14):2009–2017, 2014.

[195] Freyhult E, Moulton V, Clote P. RNAbor: a web server for RNA structural neighbors. *Nucleic Acids Research*, 35(2):W305–W309, 2009.

[196] Saffarian A, Giraud M, deMonte A, Touzet H. RNA locally optimal secondary structures. *Comput. Biol*, 19:1120–1133, 2012.

[197] Clote P. Combinatorics of saturated secondary structures of RNA. *J. Comput. Biol*, 13(9):1640–57, 2006.

[198] Fusy E, Clote P. Combinatorics of locally optimal RNA secondary structures. *J. Math. Biol*, 68(1-2):341–75, 2014.

[199] Garnier J, Kallel L. Efficiency of local search with multiple local optima. *SIAM J. Discrete Math*, 15:122–141, 2002.

[200] Sahoo S, Albrecht AA. Approximating the set of local minima in partial RNA folding landscapes. *Bioinformatics*, 28:523–530, 2012.

[201] Lois G, Blawzdziewicz J, Corey S, O'Hern CS. Protein folding on rugged energy landscapes: conformational diffusion on fractal networks. *Phys. Rev*, E81, 2010.

[202] Hawkins D. Biomeasurement. *Oxford University Press*, 2009.

[203] Watters KE, Strobel EJ, Yu AM, Lis JT, Lucks JB. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nat. Struc. & Mol. Biol*, 23:1124–31, 2016.

[204] Geary CW, Rothemund PWK, Anderson ES. A single-stranded architecture for cotranscriptional folding of RNA nanostructures. *Science*, 6198(234):799–804, 2014.

[205] Sparvath SL, Geary CW, Anderson ES. Computer-Aided Design of RNA Origami Structures. *Methods Mol. Biol*, 1500:51–80, 2017.

[206] Fallmann J, Sedlyarov V, Tanzer A, Kovarik P, Hofacker IL. AREsite2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements. *NAR*, 44:D90–D95.

[207] Yu-Cheng T. Yang et al. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 16(51), 2015.

Appendix A

Software Development and Implementation

- The `MSbind` tool was implemented in PERL and uses the *Vienna RNA package* library.
- The `RNAStrucTar` tool was implemented in C++ and PERL and uses the *Vienna RNA package*.
- The descent procedures in Chapter 6 are modified versions of those implemented in the `RNAlocmin` tool.
- The approximation heuristic was implemented in C.

  Given below is the PERL source code for the `MSbind`, source code for the other tools are available on request. Visit http://kks.inf.kcl.ac.uk for contact information.

```
#!/usr/bin/perl

# MSbind
# Calculate features of metastable microRNA target sites,
# This script requires installation of the Vienna RNA package available at:
# http://www.tbi.univie.ac.at/~ronny/RNA/index.html [Hofacker et al]
#
# Input Flags:
# -f : MSbind accepts as input a file of RNA secondary structures in dot bracket notation as
#      produced by the barrier tool from the Vienna RNA package
# -s : Target start position
# -e : Target out position
# -o : Output file
# -c : (Optional) Number of local minima to read from input file
# Usage Example:
# ./MSbind -f infile -s 224 -e 232 -o outfile -c 100

use strict;
use warnings;
use Getopt::Long;
use Storable qw(dclone);
use RNA;

##########################
#Functions
##########################
sub usage();
sub trim($);
sub getSequence();

########################
#Global variables
########################
my $start = time;
my $fname;
my $startPos;
my $endPos;
my $seq;
my $seqLength;
my $totalMinima;
my @fileInput;
my @seqArr;
my $fileOut;
my @outputList;
my @mfeStrucArr;
my @mfeTargetSite;
my $targetLength;
my $targetSeq;
my $minimaEqualToMFE = 0;
my $minimaEqualTotalPairs = 0;
my $minimaLessTotalPairs = 0;
my $minimaGreaterTotalPairs = 0;
my $lessPairingsBarrierSum = 0;
my $equalPairingBarrierSum = 0;
my $minimaLessTargetEnergy = 0;
my $minimaEqualTargetEnergy = 0;
my @minimaEqualPairings;
my @minimaLessPairings;
my @minimaGreaterPairings;
my $allMinimaTargetSum = 0;
my $maxLM;
my @allLocalMinima;
my (@inputFile, @outputFile);

#Check correct number of arguments provided
usage() if (@ARGV < 7 or
    !GetOptions('f=s' => \@inputFile, 's=i' => \$startPos, 'e=i' => \$endPos,
                'o=s' => \@outputFile, 'c:i' => \$maxLM));

    #Get inputs
    if(-e "@inputFile"){
```

```
72          open(FILE, "@inputFile");
73
74          if(defined($maxLM)){ #(Optional input) if defined input n minima
75              if($maxLM < 2){
76                  print("Concentration $maxLM must be > 1");
77              }
78              else {
79                  my $l = 0;
80                  foreach my $line (<FILE>) {
81                      chomp($line);
82                      push(@fileInput, $line);
83
84                      if($l == $maxLM){
85                          last;
86                      }
87                      $l++;
88                  }
89              }
90          }
91          else {
92              #Read all minima
93              @fileInput = <FILE>;
94              $maxLM = scalar(@fileInput)-1;
95          }
96      }
97      else {
98          print("Input file (@inputFile) does not exist!\n");
99          exit(0);
100     }
101
102     getSequence();
103     if($startPos > $endPos or $startPos < 0 or $endPos > $seqLength or
104         $startPos == $endPos){
105         print("Target site start and end position invalid\n");
106         exit(0);
107     }
108
109     ################################################
110     # Get MFE Structure, always line 1 of input
111     ################################################
112     my $mfeOpening = 0;
113     my $mfeClosing = 0;
114     my $mfePairs = 0;
115     my $mfeTotalPairs = 0;
116
117     my $mfeStructure = $fileInput[1];
118     my @tempArr = split(" ", $mfeStructure); #Split on space
119     @mfeStrucArr = split(//, $tempArr[1]);   #Get structure
120     my $mfeEnergy = $tempArr[2];             #Structure energy
121     my $mfeBarrier = $tempArr[4];            #Barrier
122
123     $targetLength = $endPos - ($startPos - 1); #Number of nucleotides at target
124     my @arrCopy = @{ dclone \@mfeStrucArr };   #Extract the target site
125     @mfeTargetSite = splice(@arrCopy, $startPos-1, $targetLength);
126
127     ################################################################
128     #Create pairing position array
129     ################################################################
130     my @basePairStack;
131     my @mfeNumArr;
132
133     for (my $pos = 0; $pos < $seqLength; $pos++){
134         if ($mfeStrucArr[$pos] eq "("){
135             push(@basePairStack, $pos);
136         }
137         elsif ($mfeStrucArr[$pos] eq ")" && @basePairStack){
138             my $y = pop(@basePairStack);
139             $mfeNumArr[$pos] = $y;
140             $mfeNumArr[$y] = $pos;
141         }
142         else
```

```
143         {
144             $mfeNumArr[$pos] = "-1";
145         }
146     }
147     @basePairStack = ();
148     ##############################################################################
149     #Get target site pairings and positions then delete pairings from structure
150     ##############################################################################
151     my $mfeOpenPosPairStr = "";
152     my $mfeClosePosPairStr = "";
153     my $mfeBasePairPosStr = "";
154 
155     for(my $pos = $startPos-1; $pos < $endPos; $pos++){
156 
157             if ($mfeStrucArr[$pos] eq "("){
158 
159                 if(scalar($mfeNumArr[$pos]) >= scalar($endPos)){
160                     my $x = $pos+1;
161                     my $y = $mfeNumArr[$pos]+1;
162                     $mfeOpenPosPairStr = $mfeOpenPosPairStr."($seqArr[$pos],
163                                         $seqArr[$mfeNumArr[$pos]])[$x,$y], ";
164                     $mfeOpening++;
165                 }
166                 elsif(scalar($mfeNumArr[$pos]) < scalar($endPos)){
167                     my $x = $pos+1;
168                     my $y = $mfeNumArr[$pos]+1;
169                     $mfeBasePairPosStr = $mfeBasePairPosStr."($seqArr[$pos],
170                                         $seqArr[$mfeNumArr[$pos]])[$x,$y], ";
171                     $mfePairs++;
172                 }
173                 $mfeStrucArr[$pos] = ".";
174                 $mfeStrucArr[$mfeNumArr[$pos]] = ".";
175             }
176             elsif ($mfeStrucArr[$pos] eq ")"){
177                 my $x = $pos+1;
178                 my $y = $mfeNumArr[$pos]+1;
179                 $mfeClosePosPairStr = "($seqArr[$mfeNumArr[$pos]],$seqArr[$pos])
180                                         [$y,$x], ".$mfeClosePosPairStr;
181                 $mfeClosing++;
182 
183                 $mfeStrucArr[$pos] = ".";
184                 $mfeStrucArr[$mfeNumArr[$pos]] = ".";
185             }
186     }
187 
188     $mfeTotalPairs = $mfeOpening + $mfeClosing + $mfePairs; #Total MFE pairings
189     my $strucNoBindingSite = "@mfeStrucArr"; #Structure without target site pairings
190     $strucNoBindingSite =~ s/(.)\s/$1/seg;
191 
192     for(my $i = $startPos-1; $i < $endPos; $i++){ #Target nucleotides
193         $targetSeq = $targetSeq."$seqArr[$i] ";
194     }
195 
196     ##########################################################################
197     # Output MFE target site
198     ##########################################################################
199     open(FILE, ">@outputFile") or die $!;
200     print FILE ("$seq\n");
201     print FILE ("Sequence Length: $seqLength\n\n");
202 
203     print FILE ("================================================================================\
204     print FILE ("- MFE Target Site Structure\n");
205     print FILE ("================================================================================\
206     print FILE ("    ($startPos) ");
207     print FILE (" $targetSeq");
208     print FILE (" ($endPos)\n");
209     print FILE ("           @mfeTargetSite\n\n");
210 
211     #Calculate energy of target site
212     my $energyStrucBefore = RNA::energy_of_struct($seq, $tempArr[1]); #Original structure
213     my $energyStrucAfter = RNA::energy_of_struct($seq, $strucNoBindingSite); #without bindings
```

```
214     my $mfeTargetEnergy = scalar($energyStrucBefore) - scalar($energyStrucAfter);
215     $mfeTargetEnergy = sprintf("%.2f", $mfeTargetEnergy);
216 
217     print FILE ("   Opening Approximation: $mfeTargetEnergy\n");
218     print FILE ("   Total Base Pairings: $mfeTotalPairs\n");
219     print FILE ("   Structure Energy: $mfeEnergy\n");
220     print FILE ("   Barrier Height: $mfeBarrier\n\n");
221 
222     #Remove trailing ,
223     $mfeOpenPosPairStr = substr($mfeOpenPosPairStr, 0, length($mfeOpenPosPairStr)-2);
224     $mfeClosePosPairStr = substr($mfeClosePosPairStr, 0, length($mfeClosePosPairStr)-2);
225     $mfeBasePairPosStr = substr($mfeBasePairPosStr, 0, length($mfeBasePairPosStr)-2);
226 
227     #Print MFE open/close/full pairing strings
228     if(scalar($mfePairs) > 0){
229         print FILE ("   Base Pairs in Target Site:");
230         print FILE (" $mfeBasePairPosStr\n\n");
231     }
232     if(scalar($mfeOpening) > 0){
233         print FILE ("   Opening:");
234         print FILE (" $mfeOpenPosPairStr\n\n");
235     }
236     if(scalar($mfeClosing) > 0){
237         print FILE ("   Closing:");
238         print FILE (" $mfeClosePosPairStr\n\n");
239     }
240 
241     ##########################################################################
242     # Minima Calculation
243     ##########################################################################
244     my $line = 2;
245     $totalMinima = $maxLM; #@fileInput-2;
246 
247     while(scalar($line) <= scalar($totalMinima)){
248         my @minima = split(' ', $fileInput[$line]); #Get and split local minimum structure
249         my @minimaArr = split(//, $minima[1]);      #Split minimum structure to array
250         my $unpairedOpening = 0;                    #Count Unpaired Open
251         my $unpairedClosing = 0;                    #Count Unpaired Closing
252         my $paired = 0;                             #Count paired
253         my $totalPairs = 0;
254         my $strucEqualMFE = 0;
255 
256         #Stores minimum information, Struc ID | Target Struc | #Pairings | Approx. Energy |
257         #Full Struc Energy | Barrier | Base Pair Types and Positions
258         my @minimaDetailArr;
259 
260         ########################################################
261         #Create base pair positioning array
262         ########################################################
263         my @basePairStack;
264         my @minimaNumArr;
265         for (my $x = 0; $x < $seqLength; $x++){
266             if ($minimaArr[$x] eq "("){
267                 push(@basePairStack, $x);
268             }
269             elsif ($minimaArr[$x] eq ")" && @basePairStack){
270                 my $y = pop(@basePairStack);
271                 $minimaNumArr[$x] = $y;
272                 $minimaNumArr[$y] = $x;
273             }
274             else
275             {
276                 $minimaNumArr[$x] = "-1";
277             }
278         }
279 
280         ########################################################
281         #Get target structure and check if equal to MFE
282         ########################################################
283         @tempArr = @{ dclone \@minimaArr };
284         my @targetStruc = splice(@tempArr, $startPos - 1, $targetLength);
```

```
285
286         if("@targetStruc" eq "@mfeTargetSite"){
287             $strucEqualMFE = 1;
288         }
289
290         ##############################################################################
291         #Get target site pairings and positions then delete pairings from structure
292         ##############################################################################
293         my $minimaOpenPosPairStr = "";
294         my $minimaClosePosPairStr = "";
295         my $minimaBasePairPosStr = "";
296
297         for(my $pos = $startPos-1; $pos < $endPos; $pos++){
298             if ($minimaArr[$pos] eq "("){
299                 if(scalar($minimaNumArr[$pos]) >= scalar($endPos)){
300                     my $x = $pos+1;
301                     my $y = $minimaNumArr[$pos]+1;
302                     $minimaOpenPosPairStr = $minimaOpenPosPairStr."($seqArr[$pos],
303                                             $seqArr[$minimaNumArr[$pos]])[$x,$y], ";
304                     $unpairedOpening++;
305                 }
306                 elsif(scalar($minimaNumArr[$pos]) < scalar($endPos)){
307                     my $x = $pos+1;
308                     my $y = $minimaNumArr[$pos]+1;
309                     $minimaBasePairPosStr = $minimaBasePairPosStr."($seqArr[$pos],
310                                             $seqArr[$minimaNumArr[$pos]])[$x,$y], ";
311                     $paired++;
312                 }
313                 $minimaArr[$pos] = ".";
314                 $minimaArr[$minimaNumArr[$pos]] = ".";
315                 $minimaNumArr[$pos] = -1;
316                 $minimaNumArr[$minimaNumArr[$pos]] = -1;
317             }
318             elsif ($minimaArr[$pos] eq ")"){
319                 my $x = $pos+1;
320                 my $y = $minimaNumArr[$pos]+1;
321                 $minimaClosePosPairStr = "($seqArr[$minimaNumArr[$pos]],$seqArr[$pos])
322                                             [$y,$x], ".$minimaClosePosPairStr;
323                 $unpairedClosing++;
324                 $minimaArr[$pos] = ".";
325                 $minimaArr[$minimaNumArr[$pos]] = ".";
326                 $minimaNumArr[$pos] = -1;
327                 $minimaNumArr[$minimaNumArr[$pos]] = -1;
328             }
329         }
330         $totalPairs = $unpairedOpening + $unpairedClosing + $paired;
331
332         ######################################################################
333         #Energy Approximation
334         ######################################################################
335         $strucNoBindingSite = "@minimaArr";
336         $strucNoBindingSite =~ s/(.)\s/$1/seg;
337         $energyStrucBefore = RNA::energy_of_struct($seq, $minima[1]);
338         $energyStrucAfter = RNA::energy_of_struct($seq, $strucNoBindingSite);
339         my $minimaTargetEnergy = $energyStrucBefore - $energyStrucAfter;
340         $minimaTargetEnergy = sprintf("%.2f", $minimaTargetEnergy);
341         $allMinimaTargetSum += $minimaTargetEnergy;
342
343         #Add minimum information to list
344         push(@minimaDetailArr, $minima[0]);          #Line
345         my $targetStrucStr = "@targetStruc";
346         $targetStrucStr =~ s/(.)\s/$1/seg;
347         push(@minimaDetailArr, $targetStrucStr);     #Structure
348         push(@minimaDetailArr, $totalPairs);         #Total pairings
349         push(@minimaDetailArr, $minimaTargetEnergy);#Target Energy Approximation
350
351         my $minimaTotalEnergy = $minima[2];          #Structure energy
352         my $minimaBarrier = $minima[4];              #Barrier Height
353         push(@minimaDetailArr, $minimaTotalEnergy);
354         push(@minimaDetailArr, $minimaBarrier);
355
```

```
356         ########################################################################
357         # If Target Site Structure is Equal to MFE check base pair positions
358         ########################################################################
359         if(scalar($strucEqualMFE) == 1){
360             my $minimaBaseStr = substr($minimaBasePairPosStr, 0,
361                                 length($minimaBasePairPosStr)-2);
362             my $minimaOpenStr = substr($minimaOpenPosPairStr, 0,
363                                 length($minimaOpenPosPairStr)-2);
364             my $minimaCloseStr = substr($minimaClosePosPairStr, 0,
365                                 length($minimaClosePosPairStr)-2);
366
367             if($mfeBasePairPosStr eq $minimaBaseStr && $mfeOpenPosPairStr eq
368                         $minimaOpenStr && $mfeClosePosPairStr eq $minimaCloseStr){
369                 push(@minimaDetailArr, 1); #Minimum is identical to MFE
370                 $minimaEqualToMFE++;
371             }
372             else {
373                 push(@minimaDetailArr, 0); #Pair positions are different
374             }
375         }
376         else {
377             push(@minimaDetailArr, 0);
378
379         }
380
381         push(@minimaDetailArr, $minimaOpenPosPairStr);
382         push(@minimaDetailArr, $minimaClosePosPairStr);
383         push(@minimaDetailArr, $minimaBasePairPosStr);
384
385         ############################
386         # Compare Pairings to MFE
387         ############################
388         if(scalar($totalPairs) == scalar($mfeTotalPairs)){
389             push(@minimaEqualPairings, \@minimaDetailArr);
390             $equalPairingBarrierSum += $minimaBarrier;
391             $minimaEqualTotalPairs++;
392         }
393         elsif(scalar($totalPairs) < scalar($mfeTotalPairs)){
394             push(@minimaLessPairings, \@minimaDetailArr);
395             $minimaLessTotalPairs++;
396             $lessPairingsBarrierSum += $minimaBarrier;
397         }
398         else {
399             push(@minimaGreaterPairings, \@minimaDetailArr);
400             $minimaGreaterTotalPairs++;
401         }
402
403         ########################################
404         # Count Structures of equal/less energy
405         ########################################
406         if(abs($minimaTargetEnergy) == abs($mfeTargetEnergy)){
407             $minimaEqualTargetEnergy++;
408
409         }
410         elsif(abs($minimaTargetEnergy) < abs($mfeTargetEnergy)){
411             $minimaLessTargetEnergy++;
412         }
413
414         push(@outputList, \@minimaDetailArr);
415         $line++;
416     }
417
418     #Sort on base pairings/Change to 3 to sort on target energy
419     @minimaLessPairings = sort{@$a[2] <=> @$b[2] } @minimaLessPairings;
420     @minimaEqualPairings = sort{@$a[2] <=> @$b[2] } @minimaEqualPairings;
421     @minimaGreaterPairings = sort{@$a[2] <=> @$b[2] } @minimaGreaterPairings;
422
423     ##############################################################################################
424     # Print Minima Information
425     ##############################################################################################
426     print FILE ("=================================================================================\
```

```
427     print FILE ("- Deep Local Minima\n");
428     print FILE ("=================================================================================\
429     my $totalMinimaLessMFE = $totalMinima-1;
430     print FILE ("   ================================================================\n");
431     print FILE ("   - Total Deep Local Minima (excluding MFE): $totalMinimaLessMFE\n");
432     print FILE ("   ================================================================\n");
433     print FILE ("   - Less base pairings: $minimaLessTotalPairs\n");
434     print FILE ("   - Equal base pairings: $minimaEqualTotalPairs\n");
435     print FILE ("   - Greater base pairings: $minimaGreaterTotalPairs\n\n");
436
437     print FILE ("   ================================================================\n");
438     print FILE ("   - Target Site Approximation Energy\n");
439     print FILE ("   ================================================================\n");
440     print FILE ("   - Less Energy: $minimaLessTargetEnergy\n");
441     print FILE ("   - Equal Energy: $minimaEqualTargetEnergy\n");
442
443
444     my $sum = $minimaLessTargetEnergy + $minimaEqualTargetEnergy;
445     print FILE ("   - < or = MFE Target Site Energy: $sum\n\n");
446
447     my $allMinimaTargetAvg = sprintf("%.2f", $allMinimaTargetSum / $totalMinimaLessMFE);
448     print FILE ("   - Avg. Opening Energy of All
449                         Minima (excluding MFE): $allMinimaTargetAvg\n\n");
450
451     ########################################
452     # Less Pairings
453     ########################################
454     my @lessBindingsCount;
455     my @lessBindingsTargetAvg;
456     for(my $i = 0; $i < @minimaLessPairings; $i++){
457         my @min = @{$minimaLessPairings[$i]};
458         my $bindings = $min[2];
459
460         if(defined($lessBindingsCount[$bindings])){
461             $lessBindingsCount[$bindings] += 1;
462             $lessBindingsTargetAvg[$bindings] += $min[3];
463         }
464         else {
465             $lessBindingsCount[$bindings] = 1;
466             $lessBindingsTargetAvg[$bindings] = $min[3];
467         }
468     }
469
470     print FILE ("   ================================================================\n");
471     print FILE ("   - Minima with Less Pairings\n");
472     print FILE ("   ================================================================\n");
473     print FILE ("   - Minima with less base pairings: $minimaLessTotalPairs\n");
474     my $n = 0;
475     my $lessPairTargetSum = 0;
476     for(my $i = 0; $i < @lessBindingsCount; $i++){
477
478         if(defined($lessBindingsCount[$i])){
479             my $avg = sprintf("%.2f", $lessBindingsTargetAvg[$i] / $lessBindingsCount[$i]);
480             $lessPairTargetSum += $lessBindingsTargetAvg[$i];
481             print FILE ("   - $n Bindings: $lessBindingsCount[$i],
482                         Average Target Site: $avg\n");
483         }
484         $n++;
485     }
486
487     if(scalar($minimaLessTotalPairs) != 0){
488         my $lessTotalAvg = sprintf("%.2f", $lessPairTargetSum/$minimaLessTotalPairs);
489         print FILE ("\n - Less Pairing Average Target Site: $lessTotalAvg\n");
490     }
491     if($minimaLessTotalPairs != 0){
492         my $lessPairingAvgBarrier = sprintf("%.2f", $lessPairingsBarrierSum /
493                                     $minimaLessTotalPairs);
494         print FILE ("   - Average Barrier Height of Minima
495             with Less Base Pairings: $lessPairingAvgBarrier\n");
496     }
497
```

```
498     ########################################
499     # Equal Bindings
500     ########################################
501     print FILE ("\n");
502     print FILE ("   ================================================================\n");
503     print FILE ("   - Minima with Equal Pairings\n");
504     print FILE ("   ================================================================\n");
505
506     my $lessEnergySum = 0;
507     my $totalEnergySum = 0;
508     my $lessEnergyCount = 0;
509     for(my $i = 0; $i < @minimaEqualPairings; $i++){
510         my @min = @{$minimaEqualPairings[$i]};
511         my $minimaTargetEnergy = $min[3];
512         $totalEnergySum += $minimaTargetEnergy;
513
514         if($minimaTargetEnergy > $mfeTargetEnergy){
515             $lessEnergySum += $minimaTargetEnergy;
516             $lessEnergyCount++;
517         }
518     }
519
520     print FILE ("   - Minima with equal base pairings: $minimaEqualTotalPairs\n");
521     print FILE ("   - Identical to MFE: $minimaEqualToMFE\n");
522     my $diff = $minimaEqualTotalPairs - $minimaEqualToMFE;
523     print FILE ("   - Different from MFE: $diff\n");
524     print FILE ("   - Number that require less energy to open: $lessEnergyCount\n");
525
526     if($minimaEqualTotalPairs != 0){
527         my $equalAvgBarrier = $equalPairingBarrierSum/$minimaEqualTotalPairs;
528         print FILE ("   ------ Equal pairing LM average barrier: $equalAvgBarrier\n");
529     }
530
531     if($lessEnergyCount != 0){
532         my $lessEnergyAvg = sprintf("%.2f", $lessEnergySum/$lessEnergyCount);
533         print FILE ("   - Average energy of minima with equal bindings as MFE
534                         but less energy (more open): $lessEnergyAvg\n");
535     }
536     else {
537         print FILE ("   - No minima with equal number of bindings to the MFE
538                         with less energy.\n");
539     }
540
541     if($minimaEqualTotalPairs != 0){
542         my $equalTotalAvg = sprintf("%.2f", $totalEnergySum/$minimaEqualTotalPairs);
543         print FILE ("   - Equal Binding Average Target Site: $equalTotalAvg\n");
544     }
545
546     ########################################
547     # Less Pairings
548     ########################################
549     print FILE ("\n");
550     print FILE ("=============================================================================================
551     print FILE (" Deep Local Minima with Less Target Site Pairs\n");
552     print FILE ("=============================================================================================
553     print FILE ("Format: Local Minima Number | Target Structure | Total Base Pairs |
554                  Target Approx. Energy | Barrier Height | Full Structure Energy | Base Pairs
555                  and Position\n\n");
556
557     if($minimaLessTotalPairs > 0){
558         printf FILE ("%-8s%-10s", "", "$targetSeq\n");
559     }
560
561     foreach my $minimum (@minimaLessPairings){
562         my $minLabel = @$minimum[0];
563         my $struc = @$minimum[1];
564         my $totalPairs = @$minimum[2];
565         my $targetSiteEnergy = @$minimum[3];
566         my $energy = @$minimum[4];
567         my $barrier = @$minimum[5];
568         my $equalToMFE = @$minimum[6];
```

```
569         my $openBasePairPos = @$minimum[7];
570         my $closeBasePairPos = @$minimum[8];
571         my $fullBasePairPos = @$minimum[9];
572
573
574         #Remove trailing , from base pairing/position strings
575         $openBasePairPos = substr($openBasePairPos, 0, length($openBasePairPos)-2);
576         $closeBasePairPos = substr($closeBasePairPos, 0, length($closeBasePairPos)-2);
577         $fullBasePairPos = substr($fullBasePairPos, 0, length($fullBasePairPos)-2);
578
579         my $format = "%-8s%-10s%-4s%-4s%-6s%-2s%-4s%-2s%-4s";
580         printf FILE ("$format", "$minLabel.", "$struc", "", "$totalPairs", "$targetSiteEnergy",
581                      "", "$barrier", "", "$energy   ");
582
583         if($openBasePairPos ne ""){
584             printf FILE ("  Opening: ");
585             printf FILE ("$openBasePairPos");
586         }
587         if($closeBasePairPos ne ""){
588             print FILE ("   Closing: ");
589             print FILE ("$closeBasePairPos");
590         }
591         if($fullBasePairPos ne ""){
592             print FILE ("   Pairing: ");
593             print FILE ("$fullBasePairPos");
594         }
595         print FILE ("\n");
596     }
597
598     ########################################
599     # Equal Pairings
600     ########################################
601     print FILE ("\n================================================================================\n");
602     print FILE (" Deep Local Minima with Equal Number of Target Site Pairs\n");
603     print FILE ("================================================================================\n");
604     if($minimaEqualTotalPairs > 0){
605         printf FILE ("%-8s%-10s", "", "$targetSeq\n");
606     }
607     foreach my $minimum (@minimaEqualPairings){
608         my $minLabel = @$minimum[0];
609         my $struc = @$minimum[1];
610         my $totalPairs = @$minimum[2];
611         my $targetSiteEnergy = @$minimum[3];
612         my $energy = @$minimum[4];
613         my $barrier = @$minimum[5];
614         my $equalToMFE = @$minimum[6];
615         my $openBasePairPos = @$minimum[7];
616         my $closeBasePairPos = @$minimum[8];
617         my $fullBasePairPos = @$minimum[9];
618
619         #Remove trailing , from base pairing/position strings
620         $openBasePairPos = substr($openBasePairPos, 0, length($openBasePairPos)-2);
621         $closeBasePairPos = substr($closeBasePairPos, 0, length($closeBasePairPos)-2);
622         $fullBasePairPos = substr($fullBasePairPos, 0, length($fullBasePairPos)-2);
623
624         my $format = "%-8s%-10s%-4s%-4s%-6s%-2s%-4s%-2s%-4s";
625         printf FILE ("$format", "$minLabel.", "$struc", "", "$totalPairs",
626                      "$targetSiteEnergy", "", "$barrier", "", "$energy   ");
627         if(scalar($equalToMFE) == 1){
628             print FILE ("   Identical to MFE ");
629         }
630         if($openBasePairPos ne ""){
631             printf FILE ("  Opening: ");
632             printf FILE ("$openBasePairPos");
633         }
634         if($closeBasePairPos ne ""){
635             print FILE ("   Closing: ");
636             print FILE ("$closeBasePairPos");
637         }
638         if($fullBasePairPos ne ""){
639             print FILE ("   Pairing: ");
```

```
640             print FILE ("$fullBasePairPos");
641         }
642         print FILE ("\n");
643     }
644
645     ########################################
646     # Greater Pairings
647     ########################################
648     print FILE ("\n================================================================================================
649     print FILE (" Deep Local Minima with a Greater Number of Target Site Pairs\n");
650     print FILE ("================================================================================================
651     if($minimaGreaterTotalPairs > 0){
652         printf FILE ("%-8s%-10s", "", "$targetSeq\n");
653     }
654
655     foreach my $minimum (@minimaGreaterPairings){
656         my $minLabel = @$minimum[0];
657         my $struc = @$minimum[1];
658         my $totalPairs = @$minimum[2];
659         my $targetSiteEnergy = @$minimum[3];
660         my $energy = @$minimum[4];
661         my $barrier = @$minimum[5];
662         my $equalToMFE = @$minimum[6];
663         my $openBasePairPos = @$minimum[7];
664         my $closeBasePairPos = @$minimum[8];
665         my $fullBasePairPos = @$minimum[9];
666
667         #Remove trailing , from base pairing/position strings
668         $openBasePairPos = substr($openBasePairPos, 0, length($openBasePairPos)-2);
669         $closeBasePairPos = substr($closeBasePairPos, 0, length($closeBasePairPos)-2);
670         $fullBasePairPos = substr($fullBasePairPos, 0, length($fullBasePairPos)-2);
671
672         my $format = "%-8s%-10s%-4s%-4s%-6s%-2s%-4s%-2s%-4s";
673         printf FILE ("$format", "$minLabel.", "$struc", "", "$totalPairs",
674                      "$targetSiteEnergy", "", "$barrier", "", "$energy   ");
675
676         if($openBasePairPos ne ""){
677             printf FILE ("  Opening: ");
678             printf FILE ("$openBasePairPos");
679         }
680         if($closeBasePairPos ne ""){
681             print FILE ("   Closing: ");
682             print FILE ("$closeBasePairPos");
683         }
684         if($fullBasePairPos ne ""){
685             print FILE ("   Pairing: ");
686             print FILE ("$fullBasePairPos");
687         }
688         print FILE ("\n");
689     }
690
691     close(FILE);
692
693     my $runtime = time - $start;
694     printf("\n\nTotal running time: %02d:%02d:%02d\n\n", int($runtime / 3600),
695             int(($runtime % 3600) / 60), int($runtime % 60));
696
697
698 ######################################################
699 # Print usage information
700 ######################################################
701 sub usage(){
702     print("\n\n");
703     print "Usage:\n
704     -f file.out - List of Local Minima Structures \r
705     -s 22 - Target Site Start Position \r
706     -e 32 - Target Site End Position \r
707     -o file_name - Output File Name \r
708     -c 100 (Optional) - Concentration, number of lowest local minima to test.\n\n";
709     print "Example:\n.\/MSbind.pl -f minima.out -s 10 -e 20 -o fileOutName.out -c 100\n\n";
710     exit(0);
```

```
711 }
712
713 ##########################################
714 # Remove leading and trailing whitespace
715 ##########################################
716 sub trim($)
717 {
718     my $str = shift;
719     $str =~ s/^\s+//;
720     $str =~ s/\s+$//;
721     return $str;
722 }
723
724 ##############################
725 # Get sequence of the input
726 ##############################
727 sub getSequence(){
728     $seq = $fileInput[0];
729     chomp($seq);
730     $seq = trim($seq);
731
732     #Split sequence
733     @seqArr = split(//, $seq);
734     $seqLength = scalar(@seqArr);
735 }
736
737 #!/usr/bin/perl
738
739 #
740 # This script combines the Vienna RNAsubopt and Barrier tools to generate
741 # local minima secondary structures. Vienna RNA package, Hofacker et al.
742 # [http://www.tbi.univie.ac.at/~ronny/RNA/index.html]
743 #
744 # This script requires the executables of RNAsubopt and Barrier tools to
745 # be in the same directory as this script.
746 #
747 #
748
749 use strict;
750 use warnings;
751 use Regexp::Common;
752 use File::Temp qw(tempfile);
753
754 my $seq;
755 my $energy;
756 my $fileInName;
757 my $outputPrefix;
758 my $outputTree = "0";
759 my $barrier = 0;
760 my $seqIsDNA = 0;
761 my $seqIsRNA = 0;
762 my $totalStructures = 0;
763 my $seqStr;
764 my $tempfile;
765 my $startTime = time;
766
767 sub countSuboptStruc();
768 sub countBarrierStruc();
769 sub validSeq($);
770 sub readSeq();
771 sub printUsage();
772 sub printHelp();
773
774 if ($#ARGV == -1){ #screen mode
775     print("\n--------------- Deep Minima Filter ---------------\n");
776     print(" 1. Generate secondary structures");
777
778     #do until a valid sequence is provided
779     until($seqIsDNA || $seqIsRNA){
780         print("\nInput the sequence or the filename of the sequence (<filename.fa)\n");
781         my $in = <STDIN>; #Input seq directly or as input file
```

```
782         chomp($in);
783
784         #Check if user input is a file or direct sequence input
785         if (substr($in, 0, 1) eq "<"){ #If input begins < then File
786
787             $fileInName = substr($in, 1, length($in)); #Remove <
788             if (-e $fileInName){                       #check file exists
789                 open FILE, $fileInName or die $!;
790                 $seqStr = <FILE>;
791                 chomp($seqStr);
792                 validSeq($seqStr)
793             }
794             else {
795                 print("\nFile does not exist\n");
796                 next;
797             }
798         }
799         else {
800             $seqStr = $in;
801             chomp($seqStr);
802             $seqStr =~ s/\s*$//;
803             validSeq($seqStr);
804
805             #Create a temporary sequence file, cannot pass string directly to RNAsubopt
806             $tempfile = new File::Temp();
807
808             print $tempfile ("$seqStr");
809             $fileInName = $tempfile;
810         }
811     }
812
813     #Energy
814     $energy = "";
815     until ($energy =~ m{^$RE{num}{real}$}){ #Check is real
816         print("\nInput energy range: ");
817         $energy = <STDIN>;
818         chomp($energy);
819         if ($energy !~ m{^$RE{num}{real}$}){
820             print "Energy value $energy is not valid.";
821         }
822     }
823
824     #Get output file prefix
825     print("Input prefix for output files: ");
826     $outputPrefix = <STDIN>;
827     chomp($outputPrefix);
828
829     my $suboptCmd;
830     #RNAsubopt
831     print("\nRNAsubopt Running...\n");
832     $suboptCmd = `./RNAsubopt --noLP -s -e $energy <$fileInName
833                 >$outputPrefix\_RNAsubopt\_offset\_$energy.out`;
834
835     system($suboptCmd);
836     #Subopt runtime
837     my $runtime = time - $startTime;
838     printf("\nRNAsubopt Complete, runtime: %02d:%02d:%02d\n\n", int($runtime / 3600),
839             int(($runtime % 3600) / 60), int($runtime % 60));
840
841     if ($tempfile){
842         $tempfile = new File::Temp( UNLINK => 0 ); #Delete temp file
843     }
844
845     countSuboptStruc(); #Count the number of structures generated
846
847     #Barrier
848     my $ans = "";
849     until(uc($ans) eq "Y" || uc($ans) eq "N"){
850         print("Would you like to run barrier to obtain deep local minima? Type Y or N\n");
851         $ans = <STDIN>;
852         chomp($ans);
```

```
853         $ans = uc($ans);
854     }
855
856     if(uc($ans) eq "N"){
857         exit 1;
858     }
859
860     print("\n2. Barrier - Deep Local Minima\n");
861     print("Enter a barrier height\n");
862
863     my $barrier = <STDIN>;
864     chomp($barrier);
865     until ($barrier =~ m{^$RE{num}{real}$}){ #Check is real
866         print("\nInput barrier height: ");
867         $barrier = <STDIN>;
868         chomp($barrier);
869         if ($barrier !~ m{^$RE{num}{real}$}){
870             print "Barrier height $barrier is not valid.";
871         }
872     }
873
874     my $answer = "";
875     until (uc($answer) eq "Y" || uc($answer) eq "N"){
876         print("Output tree diagram? Type Y or N\n");
877         $answer = <STDIN>;
878         chomp($answer);
879     }
880     if (uc($answer) eq "Y") { $outputTree = 1; }
881
882     printf("Barrier Running...\n");
883     $startTime = time;
884
885     # Barrier Options
886     #   -q, --quiet, --silent | No postscript
887     #   -v verbose
888     #   --bsize | Print the size of each basin
889     #   --max num | Compute only the lowest num local minima
890     #   --minh delta | Print only minima with energy barrier greater than delta
891     #   --saddle | Print the saddle point conformations in output
892     #   --rates | Computes the rates between macro states (basins) for use with treekin
893     #   -p l1=l2 | Compute a minimal barrier path between minima l1 and l2."
894
895     if ($outputTree eq "0"){ #-q no tree output
896         my @return = `./barriers -G RNA-noLP -q --max $totalStructures --minh $barrier
897             <$outputPrefix\_RNAsubopt\_offset\_$energy.out >$outputPrefix\_Barrier\_
898             $barrier\_offset\_$energy.out 2>&0`;
899     }
900     else {
901         my @return = `./barriers -G RNA-noLP --max $totalStructures --minh $barrier
902             <$outputPrefix\_RNAsubopt\_offset\_$energy.out >$outputPrefix\_Barrier\_
903             $barrier\_offset\_$energy.out 2>&0`;
904     }
905
906     $runtime = time - $startTime;
907     printf("\n\nBarrier Complete, runtime: %02d:%02d:%02d\n\n", int($runtime / 3600),
908             int(($runtime % 3600) / 60), int($runtime % 60));
909
910     #Count minima
911     open FILE, "$outputPrefix\_Barrier\_$barrier\_offset\_$energy.out" or die $!;
912     my @deepMinima = <FILE>;
913     close FILE;
914
915     my $numDeepMinima = 0;
916     foreach my $item(@deepMinima){
917        my @line = split(' ', $item);
918        if ($numDeepMinima == 0){
919           $numDeepMinima++;
920        }
921        else {
922           $numDeepMinima++;
923        }
```

```
924       }
925       $numDeepMinima--;
926       printf("Total number of structures with a delta value greater than $barrier
927               in file $outputPrefix\_Barrier\_$barrier.out is $numDeepMinima.\n");
928
929       exit 0;
930   }
931   elsif ($#ARGV == 0 && (($ARGV[0] eq "-help") || ($ARGV[0] eq "help"))){
932       printHelp();
933   }
934   ######################################################
935   # Command Mode
936   ######################################################
937   elsif (($#ARGV == 7) || ($#ARGV == 9)){ # (0, 1) (2, 3) (4, 5) (6, 7) Optional: (8, 9)
938                                           # -f    -e      -o      -b   -t
939
940       ################################################
941       # Check arguments, allow for any order
942       ################################################
943       if ($ARGV[0] eq "-f" && length($ARGV[0]) == 2){
944           $fileInName = $ARGV[1];
945           chomp($fileInName);
946       }
947       elsif ($ARGV[0] eq "-e" && length($ARGV[0]) == 2){
948           $energy = $ARGV[1];
949           chomp($energy);
950       }
951       elsif ($ARGV[0] eq "-b" && length($ARGV[0]) == 2){
952           $barrier = $ARGV[1];
953           chomp($barrier);
954       }
955       elsif ($ARGV[0] eq "-o" && length($ARGV[0]) == 2){
956           $outputPrefix = $ARGV[1];
957           chomp($outputPrefix);
958       }
959       elsif ($ARGV[0] eq "-t" && length($ARGV[0]) == 2){
960           if ($ARGV[1] eq "1"){
961               $outputTree = "1"; #Output tree
962           }
963       }
964       else {
965           print "\nArgument 1: \"$ARGV[0] $ARGV[1]\" is not a valid argument.\n";
966           printUsage();
967           exit 1;
968       }
969       ################################################
970       if ($ARGV[2] eq "-f" && length($ARGV[2]) == 2){
971           $fileInName = $ARGV[3];
972           chomp($fileInName);
973       }
974       elsif ($ARGV[2] eq "-e" && length($ARGV[2]) == 2){
975           $energy = $ARGV[3];
976           chomp($energy);
977       }
978       elsif ($ARGV[2] eq "-b" && length($ARGV[2]) == 2){
979           $barrier = $ARGV[3];
980           chomp($barrier);
981       }
982       elsif ($ARGV[2] eq "-o" && length($ARGV[2]) == 2){
983           $outputPrefix = $ARGV[3];
984           chomp($outputPrefix);
985       }
986       elsif ($ARGV[2] eq "-t" && length($ARGV[2]) == 2){
987           if ($ARGV[3] eq "1"){
988               $outputTree = "1";
989           }
990       }
991       else {
992           print "\nArgument 2: \"$ARGV[2] $ARGV[3]\" is not a valid argument.\n";
993           printUsage();
994           exit 1;
```

```
995       }
996       ################################################
997       if ($ARGV[4] eq "-f" && length($ARGV[4]) == 2){
998           $fileInName = $ARGV[5];
999           chomp($fileInName);
1000      }
1001      elsif ($ARGV[4] eq "-e" && length($ARGV[4]) == 2){
1002          $energy = $ARGV[5];
1003          chomp($energy);
1004      }
1005      elsif ($ARGV[4] eq "-b" && length($ARGV[4]) == 2){
1006          $barrier = $ARGV[5];
1007          chomp($barrier);
1008      }
1009      elsif ($ARGV[4] eq "-o" && length($ARGV[4]) == 2){
1010          $outputPrefix = $ARGV[5];
1011          chomp($outputPrefix);
1012      }
1013      elsif ($ARGV[4] eq "-t" && length($ARGV[4]) == 2){
1014          if ($ARGV[5] eq "1"){
1015              $outputTree = "1";
1016          }
1017      }
1018      else {
1019          print "\nArgument 3: \"$ARGV[4] $ARGV[5]\" is not a valid argument.\n";
1020          printUsage();
1021          exit 1;
1022      }
1023      ################################################
1024      if ($ARGV[6] eq "-f" && length($ARGV[6]) == 2){
1025          $fileInName = $ARGV[7];
1026          chomp($fileInName);
1027      }
1028      elsif ($ARGV[6] eq "-e" && length($ARGV[6]) == 2){
1029          $energy = $ARGV[7];
1030          chomp($energy);
1031      }
1032      elsif ($ARGV[6] eq "-b" && length($ARGV[6]) == 2){
1033          $barrier = $ARGV[7];
1034          chomp($barrier);
1035      }
1036      elsif ($ARGV[6] eq "-o" && length($ARGV[6]) == 2){
1037          $outputPrefix = $ARGV[7];
1038          chomp($outputPrefix);
1039      }
1040      elsif ($ARGV[6] eq "-t" && length($ARGV[6]) == 2){
1041          if ($ARGV[7] eq "1"){
1042              $outputTree = "1";
1043          }
1044      }
1045      else {
1046          print "\nArgument 4: \"$ARGV[6] $ARGV[7]\" is not a valid argument.\n";
1047          printUsage();
1048          exit 1;
1049      }
1050      ################################################
1051      # Optional argument
1052      ################################################
1053      if ($ARGV[8]){
1054          if ($ARGV[8] eq "-f" && length($ARGV[8]) == 2){
1055              $fileInName = $ARGV[9];
1056              chomp($fileInName);
1057          }
1058          elsif ($ARGV[8] eq "-e" && length($ARGV[8]) == 2){
1059              $energy = $ARGV[9];
1060              chomp($energy);
1061          }
1062          elsif ($ARGV[8] eq "-b" && length($ARGV[8]) == 2){
1063              $barrier = $ARGV[9];
1064              chomp($barrier);
1065          }
```

```
1066          elsif ($ARGV[8] eq "-o" && length($ARGV[8]) == 2){
1067              $outputPrefix = $ARGV[9];
1068              chomp($outputPrefix);
1069          }
1070          elsif ($ARGV[8] eq "-t" && length($ARGV[8]) == 2){
1071              if ($ARGV[9] eq "1"){
1072                  $outputTree = "1";
1073              }
1074          }
1075          else {
1076              print "\nArgument 5: \"$ARGV[8] $ARGV[9]\" is not a valid argument.\n";
1077              printUsage();
1078              exit 1;
1079          }
1080      }
1081
1082      readSeq();
1083      my $startTime = time;
1084
1085      ######################################################
1086      # Run Vienna RNAsubopt
1087      ######################################################
1088      print("\nRNAsubopt Running...\n");
1089      my $suboptCmd;
1090      #RNAsubopt
1091      $suboptCmd = `./RNAsubopt --noLP -s -e $energy <$tempfile >$outputPrefix\_
1092                    RNAsubopt\_offset\_$energy.out`;
1093      my $runtime = time - $startTime;
1094      printf("RNAsubopt complete, runtime: %02d:%02d:%02d\n\n", int($runtime / 3600),
1095              int(($runtime % 3600) / 60), int($runtime % 60));
1096
1097      countSuboptStruc(); #Count the number of structures generated
1098      print("Barrier Running...\n");
1099      $startTime = time;
1100      ######################################################
1101      # Run Barrier
1102      ######################################################
1103      if ($outputTree eq "0"){ #-q no tree output
1104          my @return = `./barriers -G RNA-noLP -q --max $totalStructures --minh
1105              $barrier <$outputPrefix\_RNAsubopt\_offset\_$energy.out >$outputPrefix\_Barrier\_
1106              $barrier\_offset\_$energy.out 2>&0`;
1107      }
1108      else {
1109          my @return = `./barriers -G RNA-noLP --max $totalStructures --minh $barrier
1110              <$outputPrefix\_RNAsubopt\_offset\_$energy.out >$outputPrefix\_Barrier\_
1111              $barrier\_offset\_$energy.out 2>&0`;
1112      }
1113      $runtime = time - $startTime;
1114      printf("Barrier complete, runtime: %02d:%02d:%02d\n\n", int($runtime / 3600),
1115              int(($runtime % 3600) / 60), int($runtime % 60));
1116
1117      countBarrierStruc();
1118  }
1119  ##########################################################################
1120  # Help Command
1121  ##########################################################################
1122  else
1123  {
1124      printUsage();
1125  }
1126
1127  ##########################################################################
1128  # Count number of suboptimal structures generated from RNAsubopt
1129  ##########################################################################
1130  sub countSuboptStruc(){
1131
1132      open(FILE, "$outputPrefix\_RNAsubopt_offset_$energy.out") or
1133          die "Cannot open '$outputPrefix\_RNAsubopt\_offset\_$energy.out': $!";
1134      while(sysread FILE, my $buffer, 4096){
1135        $totalStructures += ($buffer =~ tr/\n//);
1136      }
```

```
1137      close FILE;
1138      $totalStructures--; #First line of file is the sequence
1139      printf("   Total number of structures within energy offset: $energy\n
1140          = $totalStructures.\n\n");
1141  }
1142
1143  ########################################
1144  # Read Barrier Output
1145  ########################################
1146  sub countBarrierStruc(){
1147
1148      open FILE, "$outputPrefix\_Barrier\_$barrier\_offset\_$energy.out" or die $!;
1149      my @deepMinima = <FILE>;
1150      close FILE;
1151
1152      my $numDeepMinima = 0;
1153      foreach my $item(@deepMinima){
1154         my @line = split(' ', $item);
1155         if ($numDeepMinima == 0){
1156            $numDeepMinima++;
1157         }
1158         else {
1159            $numDeepMinima++;
1160         }
1161      }
1162      $numDeepMinima--;
1163      printf("   Total number of structures with a delta value
1164          greater than $barrier\n    = $numDeepMinima.\n\n");
1165  }
1166
1167  ##################################################
1168  # Check if a string is a valid DNA/RNA sequence
1169  ##################################################
1170  sub validSeq($){
1171
1172      if ($seqStr =~ m/([^AUTCG])/i){
1173          my @seqArr = split(//, $seqStr);
1174          my $pos = 1;
1175          foreach my $base (@seqArr){
1176              if ($base eq "$1"){
1177                  print("\nNot a valid DNA/RNA sequence!\nSequence contains
1178                      the character: $1 at position $pos\n");
1179              }
1180              else
1181              {
1182                  $pos++;
1183              }
1184          }
1185      }
1186      elsif ($seqStr =~ m/[U]/i){ #Valid RNA
1187          $seqIsRNA = 1;
1188          $seq = uc($seqStr);
1189      }
1190      elsif ($seqStr =~ m/[T]/i){ #Valid DNA
1191          $seqIsDNA = 1;
1192          $seq = uc($seqStr);
1193      }
1194  }
1195
1196  ##########################################################################
1197  # Check sequence
1198  ##########################################################################
1199  sub readSeq(){
1200
1201      if (-e $fileInName){ #Check sequence file exists
1202          open(FILE, $fileInName);
1203          my @seq = <FILE>;
1204
1205          #Check if fasta format
1206          if ($seq[0] && ($seq[0] ne "") && (substr($seq[0], 0, 1)
1207              eq ">" || substr($seq[0], 0, 2) eq " >" || substr($seq[0], 0, 2) eq " > ")){
```

```
1208            if ($seq[1]){ #Fasta format sequence on line 1
1209                $seqStr = $seq[1];
1210                chomp($seqStr);
1211                if ($seqStr eq ""){ #Empty line with \n
1212                    print "No sequence in fasta file: $fileInName\n";
1213                    exit 0;
1214                }
1215                validSeq($seqStr); #Valid sequence
1216
1217                $tempfile = new File::Temp();
1218                print $tempfile ("$seqStr");
1219            }
1220            else
1221            {
1222                print "No sequence in fasta file: $fileInName\n";
1223                exit 1;
1224            }
1225        }
1226        else #Sequence in plain text file, assume on line 0
1227        {
1228            if ($seq[0] && $seq[0] ne ""){
1229                print("Not FASTA\n");
1230                exit(0);
1231                $seqStr = $seq[0];
1232                chomp($seqStr);
1233                validSeq($seqStr);
1234            }
1235            else
1236            {
1237                print "\nSequence file is empty\nEnsure that the sequence
1238                    is on the first line or use fasta format\n.";
1239                exit 1;
1240            }
1241        }
1242    }
1243    else {
1244        print("\nFile ($fileInName) does not exist\n");
1245        exit 0;
1246    }
1247 }
1248
1249 ########################################################
1250 # Print output
1251 ########################################################
1252 sub printUsage(){
1253
1254     print "You can run this script providing no arguments or
1255         you must provide the following arguments:\n\n";
1256     print " Usage:
1257     -f file.fa - RNA/DNA Sequence File in Fasta format \r
1258     -e 2.84 - Subopt Energy Offset \r
1259     -o example1 - Output filename prefix \r
1260     -b 0.5 - Barrier height \r
1261     -t 1 - (Optional) Output Barrier Tree\n\n";
1262     print "Example run:\n./ViennaFilter.pl -f seqFileName.fa -e 8.50 -o MRPL9 -b 0.5\n\n";
1263 }
1264
1265 ########################################################
1266 # Print output
1267 ########################################################
1268 sub printHelp(){
1269     printUsage();
1270 }
1271
```

Appendix B

MicroRNA Predictions

LIG3; rs4796030; mir-221

A-Allele

STarMir predicts 13 seedless binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
9	77-93	No	-15	-3.297	-5.236	-9.764	0.669
10	77-96	No	-15.500	-3.383	-6.527	-8.973	0.628
11	77-98	No	-20.300	-3.388	-11.769	-8.531	0.577
12	81-107	No	-21.400	-0.513	-19.753	-1.647	0.460

```
Site 9

    5'->3'        G    AGA     CU   U
    Target    77   GAAA   CCAGU  GGG            93
                   ||||   |||||  |||
    miRNA     23   CUUU   GGUCG  UCU            1
    3'->5'             G            GUUACAUCGA


Site 10

    5'->3'        G    AGA     CU        N
    Target    77   GAAA   CCAGU  GG  GUGU      96
                   ||||   |||||  ||  ||||
    miRNA     23   CUUU   GGUCG  CU  UACA      1
    3'->5'             G       U   GU    UCGA


Site 11

    5'->3'        G    AGA     CUG       G
    Target    77   GAAA   CCAGU    GGUGUGG    98
                   ||||   |||||    |||||||
    miRNA     23   CUUU   GGUCG    UUACAUC    1
    3'->5'             G       UCUG       GA

The bindings in green here highlight positions predicted by
PITA (94-99)

Site 12

    5'->3'        A     AGUCU  GU UG GA   C   A
    Target    81   AGACC     GG  G  G  AUG AGC   107
                   |||||     ||  |  |  ||| |||
    miRNA     23   UUUGG     UC  C  U  UAC UCG   1
    3'->5'        C     G      GU UG      A   A
```

PITA predicts 2 sites with an overall score -5.98.
There is no binding site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-221	94-99	6:1:1	-15.6	-9.61	-5.98

C-Allele

STarMir predicts 14 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
9	80-93	No	-19.300	-0.131	-8.888	-10.412	0.492
10	80-96	No	-19.800	-0.176	-9.749	-10.051	0.490
11	80-88	No	-18	-0.131	-8.513	-9.487	0.549
12	80-98	No	-24.600	-0.197	-15.961	-8.639	0.441
13	80-107	No	-25.500	-0.160	-23.586	-1.914	0.406

```
Site 9

    5'->3'        A         CU   U
    Target    80   AAGCCCAGU  GGG            93
                   |||||||||  |||
    miRNA     23   UUUGGGUCG  UCU            1
    3'->5'        C              GUUACAUCGA


Site 10

    5'->3'        A         CU        N
    Target    80   AAGCCCAGU  GG  GUGU      96
                   |||||||||  ||  ||||
    miRNA     23   UUUGGGUCG  CU  UACA      1
    3'->5'        C         U   GU    UCGA


Site 11

    5'->3'        A         C
    Target    80   AAGCCCAGU               88
                   |||||||||
    miRNA     23   UUUGGGUCG               1
    3'->5'        C         UCUGUUACAUCGA


Site 12

    5'->3'        A         CUG       G
    Target    80   AAGCCCAGU    GGUGUGG    98
                   |||||||||    |||||||
    miRNA     23   UUUGGGUCG    UUACAUC    1
    3'->5'        C         UCUG       GA


The bindings in green highlight positions predicted by PITA
(94-99)
```

PITA predicts 2 sites with an overall score -7.45.
There is no binding site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	hsa-miR-221	94-99	6:1:1	-19.8	-12.34	-7.45

CBR1; rs9024; mir-574-5p

G-Allele

STarMir predicts 15 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
10	121-162	No	-27.800	-5.364	-8.085	-19.715	0.527
11	121-137	No	-21.300	-3.355	-3.283	-18.017	0.528
12	121-140	No	-21.600	-3.538	-4.309	-17.291	0.565

Site 10

```
    5'->3'         A        UAAUGUAC   UAAUUGAGCAACCU         G
    Target   121    GCACUCAC         UAC               ACGCACUCA   162
                   ||||||||         |||               |||||||||
    miRNA     23    UGUGAGUG         GUG               UGUGUGAGU   1
    3'->5'         UG        U
```

The bindings in green highlight positions predicted by PITA
154-161 with a perfect 8 mer with one G-U wobble

Site 11

```
    5'->3'         A        UA         A
    Target   121    GCACUCAC    AUGUACU    137
                   ||||||||    |||||||
    miRNA     23    UGUGAGUG    UGUGUGA    1
    3'->5'         UG        UGUG       GU
```

Site 12

```
    5'->3'         A        UAAUG        A
    Target   121    GCACUCAC      UAC UACU    140
                   ||||||||      ||| ||||
    miRNA     23    UGUGAGUG      GUG GUGA    1
    3'->5'         UG        UGU     U    GU
```

PITA predicts 3 sites with an overall score -17.94.
There is no binding site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	hsa-miR-574-5p	154-161	8:0:1	-24.49	-6.54	-17.94

A-Allele

STarMir predicts 16 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
10	121-137	No	-25.300	-4.513	-2.874	-22.426	0.584
11	121-140	No	-25.500	-4.599	-3.514	-21.986	0.631
12	121-162	No	-27.800	-5.224	-5.958	-21.842	0.563
13	121-135	No	-20.100	-3.708	-3.88	-16.220	0.574

Site 10

```
    5'->3'         A        UA         A
    Target   121    GCACUCAC    AUAUACU    137
                   ||||||||    |||||||
    miRNA     23    UGUGAGUG    UGUGUGA    1
    3'->5'         UG        UGUG       GU
```

Site 11

```
    5'->3'         A        UA           A
    Target   121    GCACUCAC  AUAUAC UACU    140
                   ||||||||  |||||| ||||
    miRNA     23    UGUGAGUG  UGUGUG GUGA    1
    3'->5'         UG                U    GU
```

Site 12

```
    5'->3'         A        UAAUAUAC   UAAUUGAGCAACCU         G
    Target   121    GCACUCAC         UAC               ACGCACUCA   162
                   ||||||||         |||               |||||||||
    miRNA     23    UGUGAGUG         GUG               UGUGUGAGU   1
    3'->5'         UG        U
```

The bindings in green highlight positions predicted by PITA
154-161 with a perfect 8 mer with one G-U wobble

Site 13

```
    5'->3'         A        UA       N
    Target   121    GCACUCAC    AUAUA      135
                   ||||||||    |||||
    miRNA     23    UGUGAGUG    UGUGU      1
    3'->5'         UG        UGUG     GAGU
```

PITA predicts 4 sites with an overall score -19.21.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-574-5p	133-138	6:1:1	-19.8	-2.56	-17.23

The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-574-5p	154 - 161	8:0:1	-24.49	-5.41	-19.07

FindTar predicts 3 binding sites for each sequence. Only the A-Allele has one site covering the SNP position:

hsa-miR-574-5p	A-Allele	133-162	3' UGUG-UGA-----GUGU-GUGUGUGUGAGU 5' |:||*|||******:||***:||:|||||* 5' ATACTACTAATTGAGCAACCTACGCACTCA 3'	25.00	-22.40	excellent

HTR3E; rs56109847; miR-510-5p

G-Allele

STarMir predicts 16 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
6	50-80	73-79	-28.700	-5.505	-12.645	-16.055	0.544
7	50-76	No	-20.300	-6.324	-11.509	-8.791	0.580

```
Site 6

    5'->3'          G  CUG    GGUCUCCCCC            C
    Target    50     GA   GCCA          CUUUCCUGAGUA    80
                     ||   ||||          ||||||||||||
    miRNA     22     CU   CGGU          GAGAGGACUCAU    1
    3'->5'         CA  AA

The bindings in green highlight positions predicted by PITA
72-79 with a perfect 8 mer.

Site 7

    5'->3'          G  CUG    GGUCUCCCCC        N
    Target    50     GA   GCCA          CUUUCCUG      76
                     ||   ||||          ||||||||
    miRNA     22     CU   CGGU          GAGAGGAC      1
    3'->5'         CA  AA                      UCAU
```

PITA predicts 6 sites with an overall score -16.41.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	miR-510-5p	72-79	8:0:0	-22.5	-6.08	-16.41

FindTar predictions

Position	Structure	Loop Score	ΔG
60-80	3' CACUAACGGUGAGAGGACUCAU 5' *******||*||:|||||||| 5' GTCTCCCCC-CTTTCCTGAGTA 3'	10.00	-24.90

A-Allele

STarMir predicts 16 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
6	50-80	No	-22.500	-2.464	-8.261	-14.239	0.566

```
Site 6

    5'->3'          G  CUG    GGUCUCCCCC        A    C
    Target    50     GA   GCCA          CUUUCCU AGUA    80
                     ||   ||||          ||||||| ||||
    miRNA     22     CU   CGGU          GAGAGGA UCAU    1
    3'->5'         CA  AA                      C

The bindings in green highlight positions predicted by PITA
72-79 with a 8 mer with one mismatch
```

PITA predicts 6 sites with an overall score -12.28.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	miR-510-5p	72-79	8:1:0	-16.3	-4.01	-12.28

HLA_G; rs1063320; mir-148a-3p

C-Allele

STarMir predicts 6 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
4	221-239	No	-23.900	-4.554	-9.282	-14.618	0.492
5	221-237	No	-18.900	-4.182	-9.628	-9.272	0.475
6	230-260	No	-9.800	-3.556	-9.067	-0.733	0.528

```
STarMir Site 4

    5'->3'        U              C      G
    Target   221   CAAA  UUUGUGGU CACUGA   239
                   ||||  |||||||| ||||||
    miRNA     22   GUUU  AGACAUCA GUGACU   1
    3'->5'        U    CA         C

The bindings in green highlight positions predicted by PITA
(231-238 with one mismatch)

STarMir Site 5

    5'->3'        U              C    N
    Target   221   CAAA  UUUGUGGU CACU    237
                   ||||  |||||||| ||||
    miRNA     22   GUUU  AGACAUCA GUGA    1
    3'->5'        U    CA         C    CU


STarMir Site 6

    5'->3'           U   CCA       CUAUAACUUACUUC       A
    Target   230       GGU    CUG AG                UGUAUU    260
                      |||    ||| ||                ||||||
    miRNA     22      UCA    GAC UC                ACGUGA    1
    3'->5'        UGUU   A      A                      CU
```

PITA predicts 7 sites with an overall score -1.16.
There is only one site covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	hsa-miR-148a	231-238	8:1:0	-16.8	-15.78	-1.01

FindTar predictions

220-239	3' UGUUUCAAGACAUCACGUGACU 5' *||||**|:|||:||*|||||| 5' TCAAA--TTTGTGGTCCACTGA 3'	-21.60

G-Allele

STarMir predicts 7 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
4	221-239	232-238	-30.500	-3.238	-11.006	-19.494	0.395
5	221-235	No	-21.100	-1.928	-9.059	-12.041	0.338
6	221-233	No	-15.600	-1.501	-11.349	-4.251	0.339

```
STarMir Site 4 (seed site)

    5'->3'        U                    G
    Target   221   CAAA  UUUGUGGUGCACUGA   239
                   ||||  ||||||||||||||
    miRNA     22   GUUU  AGACAUCACGUGACU   1
    3'->5'        U    CA

The bindings in green highlight positions predicted by PITA.(
231-238 with perfect seed match)

STarMir Site 5

    5'->3'        U                 N
    Target   221   CAAA  UUUGUGGUGCA       235
                   ||||  ||||||||||
    miRNA     22   GUUU  AGACAUCACGU       1
    3'->5'        U    CA            GACU


STarMir Site 6

    5'->3'        U              N
    Target   221   CAAA  UUUGUGGUG         233
                   ||||  |||||||||
    miRNA     22   GUUU  AGACAUCAC         1
    3'->5'        U    CA          GUGACU
```

PITA predicts 7 sites with an overall score -6.49.
There is only one site covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	hsa-miR-148a	231-238	8:0:0	-23.4	-16.90	-6.49

220-239	3' UGUUUCAAGACAUCACGUGACU 5' *||||**|:|||:|||||||| 5' TCAAA--TTTGTGGTGCACTGA 3'	-28.20

PARP1; rs8679; mir-145-5p

T-Allele

STarMir predicts 25 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
22	592-611	No	-18.300	-0.927	-10.683	-7.617	0.476
23	592-614	No	-19.700	-0.927	-12.191	-7.509	0.517

```
Site 22

    5'->3'            U     UCUCCA    U    A
    Target   592       UUCCU       GGAA ACUG     611
                       |||||       |||| ||||
    miRNA     23       AAGGA       CCUU UGAC     1
    3'->5'         UCCCU     C          U    CUG


Site 23

    5'->3'            U     UCUCCA    U    A  A
    Target   592       UUCCU       GGAA ACUG AC   614
                       |||||       |||| |||| ||
    miRNA     23       AAGGA       CCUU UGAC UG   1
    3'->5'         UCCCU     C          U    C

The bindings in green highlight positions predicted by PITA
608-614 with a 6 seed match with one mismatch
```

PITA predicts 7 sites with an overall score -2.53.
There is no binding site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
T-Allele	hsa-miR-145-5p	608-614	6:1:0	-12.8	-13.30	0.50
T-Allele	hsa-miR-145-5p	615-621	6:1:1	-8.64	-12.82	4.18

C-Allele

STarMir predicts 25 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
22	592-611	No	-17.800	-0.652	-17.666	-0.134	0.330
23	592-614	No	-19.200	-0.652	-20.904	1.704	0.347

```
Site 22

    5'->3'            U     UCUCCA    C    A
    Target   592       UUCCU       GGAA ACUG     611
                       |||||       |||| ||||
    miRNA     23       AAGGA       CCUU UGAC     1
    3'->5'         UCCCU     C          U    CUG


Site 23

    5'->3'            U     UCUCCA    C    A  A
    Target   592       UUCCU       GGAA ACUG AC   614
                       |||||       |||| |||| ||
    miRNA     23       AAGGA       CCUU UGAC UG   1
    3'->5'         UCCCU     C          U    C

The bindings in green highlight positions predicted by PITA
608-614 with a 6 seed match with one mismatch
```

PITA predicts 7 sites with an overall score -2.51.
There is no binding site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	hsa-miR-145-5p	608-614	6:1:0	-12.6	-13.33	0.73
C-Allele	hsa-miR-145-5p	615-621	6:1:1	-8.64	-12.80	4.16

WFS1; rs1046322; hsa-miR-668-3p

G-Allele

STarMir predicts 75 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
25	234-258	251-257	-26.600	-4.421	-10.641	-15.959	0.498

```
Site 25

    5'->3'                C       UGACCUUUCU         U
    Target   234           CUGAGCC           GAGUGACA   258
                           |||||||           ||||||||
    miRNA     23          GGCUCGG           CUCACUGU   1
    3'->5'         CAUCACCC
```

The bindings in green highlight positions predicted by PITA 251-257 with a perfect 7 mer.

PITA predicts 15 sites with an overall score -16.47.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	hsa-miR-668-3p	251	7:0:0	-24.05	-7.57	-16.47

FindTar predictions

Position	Structure	Loop Score	ΔG
236-258	3' CAUCACCCGGCUCGG----------CUCACUGU 5' ********|:||||**********|||||||* 5' TCTCCACCCTGAGCCTGACCTTTCTGAGTGACA 3'	20.00	-25.60

A-Allele

STarMir predicts 75 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
25	234-258	No	-20.600	-1.723	-14.196	-6.404	0.467

```
Site 25

    5'->3'                C       UGACCUUUCU  A     U
    Target   234           CUGAGCC           GA UGACA   258
                           |||||||           || |||||
    miRNA     23          GGCUCGG           CU ACUGU   1
    3'->5'         CAUCACCC                    C
```

The bindings in green highlight positions predicted by PITA 251-257 with a 7 mer with one mismatch

PITA predicts 15 sites with an overall score -11.13.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-668-3p	251	7:1:0	-18.05	-6.92	-11.12

IL23R; rs10889677; let-7e

C-Allele

STarMir predicts 39 seedless binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
13	291-310	No	-24.700	0	-25.819	1.119	0.201

```
Site 13

    5'->3'          U   GC  U            U
    Target   291     UUA  CA UCUUCUGCCUCA   310
                     |||  || ||||||||||||
    miRNA     22     GAU  GU GGAGGAUGGAGU   1
    3'->5'         UU   AU  U
```

The bindings in green highlight positions predicted by PITA 303 -309 (7 mer with one G-U wobble)

PITA predicts 14 sites with an overall score -14.23.
There is one binding site covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	hsa-let-7e	303 -309	7:0:1	-19.4	-5.27	-14.12

FindTar predictions

Position	Structure	Loop Score	ΔG
287-310	3' UUGAUAUGUUGGAGGAUGGAGU 5' **:||**||**||:||:||||* 5' TTTTAGCCATTCTTCTGCCTCA 3'	20.00	-23.50

A-Allele

STarMir predicts 38 binding sites.
There is no binding site covering the SNP position.

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
13	291-308	No	-20.700	-0.045	-21.396	0.696	0.205

```
Site 13

    5'->3'          U   GC  U           A
    Target   291     UUA  CA UCUUCUGCCU     308
                     |||  || ||||||||||
    miRNA     22     GAU  GU GGAGGAUGGA    1
    3'->5'         UU   AU  U           GU
```

The bindings in green highlight positions predicted by PITA 303 -309 (7 mer with one G-U wobble and one mismatch)

PITA predicts 14 sites with an overall score -12.11.
There is one binding site covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-let-7e	303-309	7:1:1	-15.9	-5.87	-10.02

Position	Structure	Loop Score	ΔG
287-310	3' UUGAUAUGUUGGAGGAUGGAGU 5' **:||**||**||:||:|||** 5' TTTTAGCCATTCTTCTGCCTAA 3'	20.00	-20.00

RYR3 ; rs1044129 ; miR-367

A-Allele

STarMir predicts 12 seedless binding sites and one seed site.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
11	830-847	No	-16.300	-0.335	-7.633	-8.667	0.582
12	835-857	852-857	-14.800	-1.222	-11.419	-3.381	0.532

```
STarMir Site 11

    5'->3'        C   AAUACAA        C
    Target   830   UCA       UGAAGUGC       847
                   |||       ||||||||
    miRNA     22   AGU       AUUUCACG       1
    3'->5'           GGUAACG        UUAA


Site 12 (seed site)
The bindings in green highlight positions predicted by PITA
but with A-U binding at the beginning (850-857)

    5'->3'        A   AAUGAAG    CCAC       A
    Target   835   UAC       UGC     UGCAAU   857
                   |||       |||     ||||||
    miRNA     22   GUG       ACG     ACGUUA   1
    3'->5'        A   GUA       AUUUC      A
```

PITA predicts 29 binding sites with an overall score -6.97.
There is only one binding site covering the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-367	835-840	6:1:0	-1.77	-6.76	4.99

The following sites may cause the opening of the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-367	850-857	8:1:0	-8.9	-12.16	3.26
A-Allele	hsa-miR-367	840-845	6:1:1	-1.95	-9.23	7.28

G-Allele

STarMir predicts 12 seedless binding sites and one seed site.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
11	830-847	No	-16.300	-0.840	-11.825	-4.475	0.463
12	830-857	852-857	-15.300	-0.194	-19.650	4.350	0.397

```
Site 11

    5'->3'        C   AAUA  G          C
    Target   830   UCA     CA    UGAAGUGC       847
                   |||     ||    ||||||||
    miRNA     22   AGU     GU    AUUUCACG       1
    3'->5'           G      AACG       UUAA


Site 12 (seed site)
The bindings in green highlight positions predicted by PITA but
with A-U binding at the beginning (850-857)

    5'->3'        C   AAUA  GUGAAG    CCAC       A
    Target   830   UCA     CA      UGC     UGCAAU   857
                   |||     ||      |||     ||||||
    miRNA     22   AGU     GU      ACG     ACGUUA   1
    3'->5'           G      A         AUUUC      A
```

PITA predicts 29 binding sites with an overall score -6.97.
There are two binding sites covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	hsa-miR-367	838-845	8:1:1	-7.7	-11.38	3.68
G-Allele	hsa-miR-367	835-840	6:1:1	-1.77	-8.92	7.15

The following site may cause the opening of the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	hsa-miR-367	850-857	8:1:0	-8.73	-14.31	5.58

AGTR1; rs5186 ; miR-155-5p

A-Allele

STarMir predicts 21 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
2	57-90	83-89	-20.900	-0.577	-28.561	7.661	0.283
3	57-86	No	-15.700	-0.085	-27.289	11.589	0.248

```
Site 2

    5'->3'        U      CAGCACU    UACCAAAUG        C
    Target    57   CCUCUG       UCAC          AGCAUUAG   90
                   ||||||       ||||          ||||||||
    miRNA     23   GGGGAU       AGUG          UCGUAAUU   1
    3'->5'        U                  CUAA

The bindings in green highlight positions predicted by PITA
83-89 with a perfect 7 mer.

Site 3

    5'->3'        U      CAGCACU    UACCAA  G    N
    Target    57   CCUCUG       UCAC      AU AGCA      86
                   ||||||       ||||      || ||||
    miRNA     23   GGGGAU       AGUG      UA UCGU      1
    3'->5'        U                  C       A    AAUU
```

PITA predicts 15 sites with an overall score -5.40.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	miR-155-5p	83-89	7:0:0	-14.27	-9.97	-4.29

The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	miR-155-5p	100	7:1:1	-2.12	-10.04	7.92

FindTar predictions

Position	Structure	Loop Score	ΔG
67-90	3' UGGGGAUAGUGCUAAUCGUAAUU 5' ****|||*||*****|||||||* 5' TTCACTACCA-AATGAGCATTAG 3'	20.00	-15.50

C-Allele

STarMir predicts 20 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
2	57-90	No	-16.600	-2.268	-27.879	11.279	0.284
3	84-123	118-123	-14.900	-0.429	-22.035	7.135	0.475

```
Site 2

    5'->3'        U      CAGCACU    UACCAAAUG   C    C
    Target    57   CCUCUG       UCAC          AGC UUAG   90
                   ||||||       ||||          ||| ||||
    miRNA     23   GGGGAU       AGUG          UCG AAUU   1
    3'->5'        U                  CUAA        U

The bindings in green highlight positions predicted by PITA
83-89 with a 7 mer with one mismatch

Site 3

    5'->3'        A    UAG   CUUU   GA    AAGGAGAAAAU      U
    Target    84   GCCU   CUA    UCA  AUUG           GCAUUA
123
                   ||||   |||    |||  ||||           ||||||
    miRNA     23   UGGG   GAU    AGU  UAAU           CGUAAU   1
    3'->5'                          GC                    U
```

PITA predicts 15 sites with an overall score -5.19.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	miR-155-5p	83-89	7:1:0	-9.97	-6.84	-3.12

The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	miR-155-5p	100	7:1:1	-7.7	-6.78	-0.91

FGF20; rs12720208; miR-433-3p

C-Allele

STarMir predicts 45 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
9	166-187	180-186	-14.500	-2.356	-9.274	-5.226	0.426

```
Site 9

    5'->3'          U    CU  AAAUAG        C
    Target   166     UUGA  AG      AUCAUGAU   187
                     ||||  ||      ||||||||
    miRNA     22     GGCU  UC      UAGUACUA   1
    3'->5'         UGU    CC  GGG

The bindings in green highlight positions predicted by PITA
180-186 with a perfect 7 mer.
```

PITA predicts 22 sites with an overall score -11.08.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	miR-433-3p	180-186	7:0:0	-10.55	-7.11	-3.43

FindTar predictions

Position	Structure	Loop Score	ΔG
163-187	3' UGUGGCUCCUC--GG-GUAGUACUA 5' ***::||**||***:**|||||||| 5' ATTTTGACTAGAAATAGATCATGAT 3'	20.00	-12.80

T-Allele

STarMir predicts 46 binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
9	166-187	No	-12.300	-4.655	-10.123	-2.177	0.547
10	166-185	No	-8.600	-3.477	-10.088	1.488	0.526

```
Site 9

    5'->3'          U    CU  AAAUAG        C
    Target   166     UUGA  AG      AUUAUGAU   187
                     ||||  ||      ||||||||
    miRNA     22     GGCU  UC      UAGUACUA   1
    3'->5'         UGU    CC  GGG

The bindings in green highlight positions predicted by PITA
180-186 with a 7 mer with one G-U wobble
```

PITA predicts 22 sites with an overall score -11.08.
There is one binding site covering the SNP position.

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
T-Allele	miR-433-3p	180-186	7:0:1	-8.35	-6.70	-1.64

Position	Structure	Loop Score	ΔG
163-187	3' UGUGGCUCCUC--GG-GUAGUACUA 5' ***::||**||***:**||:||||| 5' ATTTTGACTAGAAATAGATTATGAT 3'	20.00	-10.60

HOXB5 ; rs9299 ; miR-7-5p

G-Allele

STarMir predicts 26 seedless binding sites and one seed site.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
1	126-145	No	-17.200	-0.428	-6.624	-10.576	0.363
2	126-154	148-154	-21.900	-0.493	-13.609	-8.291	0.389

STarMir Site 1

```
    5'->3'        U      UUUCGU         A
    Target   126   ACGAUA       UUGGUCUU      145
                  ||||||       ||||||||
    miRNA     23  UGUUGU       GAUCAGAA      1
    3'->5'                UUUAGU         GGU
```

STarMir Site 2 seed site

```
    5'->3'        U      UUUCGUUU    UUA         U
    Target   126   ACGAUA         GGUC   GGUCUUCC    154
                  ||||||         ||||   ||||||||
    miRNA     23  UGUUGU         UUAG   UCAGAAGG    1
    3'->5'                U            UGA         U
```

The bindings in green highlight positions predicted by PITA (147-154)

PITA predicts 41 binding sites with an overall score -8.39.
There is no binding site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	hsa-miR-7	148-154	7:0:0	-16.4	-9.5	-6.89

FindTar predictions

Position	Structure	Loop Score	ΔG
126-155	3' UGUUGUUUUAGUGAUCAGAAGGU 5' *******::||***:||||||||* 5' TTCGTTTGGTCTTAGGTCTTCCT 3'	25.00	-20.80

A-Allele

STarMir predicts 26 seedless binding sites and one seed site.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
1	126-154	148-154	-22.400	-1.391	-11.198	-11.202	0.466
2	126-152	No	-15.100	-1.183	-9.139	-5.961	0.463

STarMir Site 1 seed site

```
    5'->3'        U      UUUCGUUU    UUA         U
    Target   126   ACGAUA         GAUC   GGUCUUCC    154
                  ||||||         ||||   ||||||||
    miRNA     23  UGUUGU         UUAG   UCAGAAGG    1
    3'->5'                U            UGA         U
```

The bindings in green highlight positions predicted by PITA (147-154)

STarMir Site 2

```
    5'->3'        U      UUUCGUUU    U A       N
    Target   126   ACGAUA         GAUC U GGUCUU      152
                  ||||||         |||| | ||||||
    miRNA     23  UGUUGU         UUAG G UCAGAA      1
    3'->5'                U            U A       GGU
```

PITA predicts 41 binding sites with an overall score -10.25.
There is no binding site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-7	148-154	7:0:0	-16.9	-6.8	-10.09

Position	Structure	Loop Score	ΔG
126-155	3' UGUUGUUUUAGUGAUCAGAAGGU 5' *******:|||***:|||||||* 5' TTCGTTTGATCTTAGGTCTTCCT 3'	25.00	-21.30

RAD51; rs7180135; Mir-197-3p

G-Allele

STarMir predicts 22 binding sites and 2 seed sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
14	707-725	718-724	-28.600	-0.561	-14.289	-14.311	0.380
15	707-721	No	-20.700	-0.076	-14.934	-5.766	0.298
16	707-719	No	-15.100	-0.076	-15.177	0.077	0.338

```
Site 14

    5'->3'        A          CC           A
    Target   707   GCU    UGG  AAGGUGGUGAA   725
                   |||    |||  ||||||||||
    miRNA     22   CGA    ACC  UUCCACCACUU   1
    3'->5'            CCC    UC
The bindings in green here highlight positions predicted by
PITA 717-724 (perfect 8 mer seed match)


Site 15

    5'->3'        A          CC       N
    Target   707   GCU    UGG  AAGGUGG      721
                   |||    |||  |||||||
    miRNA     22   CGA    ACC  UUCCACC      1
    3'->5'            CCC    UC       ACUU


Site 16

    5'->3'        A          CC     N
    Target   707   GCU    UGG  AAGGU        719
                   |||    |||  |||||
    miRNA     22   CGA    ACC  UUCCA        1
    3'->5'            CCC    UC     CCACUU
```

PITA predicts 15 sites with an overall score -12.50.
The following are those covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
G-Allele	hsa-miR-197-3p	716-721	6:1:1	-12.21	-12.94	0.73
G-Allele	hsa-miR-197-3p	717-724	8:0:0	-22	-13.61	-8.38

FindTar predictions

Position	Structure	Loop Score	ΔG
707-725	3' CGACCCACCUCUUCCACCACUU 5' |||***|||**||||||||||* 5' GCT---TGGCCAAGGTGGTGAA 3'	20.00	-25.90

A-Allele

STarMir predicts 21 binding sites and 2 seed site.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
14	707-725	719-724	-22	-1.201	-15.003	-6.997	0.348
15	707-723	No	-18	0	-15.233	-2.767	0.279

```
Site 14

    5'->3'        A          CC   A       A
    Target   707   GCU    UGG  AAG UGGUGAA   725
                   |||    |||  ||| |||||||
    miRNA     22   CGA    ACC  UUC ACCACUU   1
    3'->5'            CCC    UC   C

The bindings in green here highlight positions predicted by
PITA 717-724 (8 mer with a mismatch)

Site 15

    5'->3'        A          CC   A     N
    Target   707   GCU    UGG  AAG UGGUG    723
                   |||    |||  ||| |||||
    miRNA     22   CGA    ACC  UUC ACCAC    1
    3'->5'            CCC    UC   C     UU
```

PITA predicts 14 sites with an overall score -12.48.
There is one binding site covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-197-3p	717-724	8:1:0	-15.4	-13.96	-1.43

Position	Structure	Loop Score	ΔG
707-725	3' CGACCCACCUCUUCCACCACUU 5' |||***|||**|||*|||||* 5' GCT---TGGCCAAGATGGTGAA 3'	20.00	-19.30

ORAI1 ; rs76753792 ; mir-519a-3p

C-Allele

STarMir predicts 4 seed sites and 32 seedless binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
2	69-88	No	-16.700	-4.322	-9.493	-7.207	0.562
3	85-102	No	-18.300	-0.019	-21.657	3.357	0.359

```
Site 2

    5'->3'        A      AC       CAGCC     A
    Target   69   GC CUC      GGA      UGCGC     88
                  || |||      |||      |||||
    miRNA    22   UG GAG      CCU      ACGUG     1
    3'->5'          U   AUUUU             AAA


Site 3

    5'->3'        U             C  G    C
    Target   85   GCGC   AGGGGG UG GCUU   102
                  ||||   |||||| || ||||
    miRNA    22   UGUG   UUUUCC AC UGAA   1
    3'->5'           AGA      U  G    A
```

The bindings in green highlight positions predicted by PITA (a 7 mer 96-102 with one mismatch and one G-U wobble)

PITA predicts 18 sites with an overall score -10.66.
There is no site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	hsa-miR-519a-3p	96-102	7:1:1	-12.7	-20.64	7.94

T-Allele

STarMir predicts 4 seed sites and 32 seedless binding sites.
The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
2	81-95	No	-15.200	0	-17.396	2.196	0.363
3	81-102	No	-17.900	-0.023	-23.521	5.621	0.381

```
Site 2

    5'->3'        A      UGC   GG  U
    Target   81    GC CUG   AGG  GC       95
                  || |||   |||  ||
    miRNA    22    UG GAU   UCC  CG       1
    3'->5'       UG  A   UU    UA  UGAAA


Site 3

    5'->3'        A     G  C      C  G    C
    Target   81   GC CU UG AGGGGG UG GCUU   102
                  || || || |||||| || ||||
    miRNA    22   UG GA AU UUUCCU AC UGAA   1
    3'->5'          U  G            G    A
```

The bindings in green highlight positions predicted by PITA (a 7 mer 96-102 with one mismatch and one G-U wobble)

PITA predicts 18 sites with an overall score -10.66.
There is no site covering the SNP position.
The following site may cause the opening of the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
T-Allele	hsa-miR-519a-3p	96-102	7:1:1	-12.8	-22.88	10.08

RAP1 ; rs6573 ; hsa-miR-196a

C-Allele

STarMir predicts 37 seedless binding sites. The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
18	348-370	No	-16.700	-3.971	-5.402	-11.298	0.623

```
STarMir Site 18

    5'->3'        A   UUUUAU     UC  C    U
    Target   348   UCU       CAUGA  CU CCUA   370
                   |||       |||||  || ||||
    miRNA     22   GGG       GUACU  GA GGAU   1
    3'->5'             UUGUU       UU  U
```

The bindings in green highlight positions predicted by PITA (364-369 a 6 seed match with one mismatch)

PITA predicts 22 binding sites with an overall score -9.02.
There is one binding site covering the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	hsa-miR-196a	364-369	6:1:0	-12.5	-5.03	-7.46

The following site may cause the opening of the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
C-Allele	hsa-miR-196a	368-373	6:1:1	-9.4	-3.22	-6.17

A-Allele

STarMir predicts 37 seedless binding sites and one seed site. The following are those covering the SNP position:

Site ID	Site Position	Seed	ΔG_{hybrid}	ΔG_{nucl}	ΔG_{open}	ΔG_{total}	Site Access
18	348-370	364-369	-21.300	-6.967	-5.051	-16.249	0.616
19	348-368	No	-17.500	-5.900	-4.889	-12.611	0.600

```
STarMir Site 18 seed site

    5'->3'        A   UUUUAU     UC       U
    Target   348   UCU       CAUGA  CUACCUA   370
                   |||       |||||  |||||||
    miRNA     22   GGG       GUACU  GAUGGAU   1
    3'->5'             UUGUU       UU
```

The bindings in green highlight positions predicted by PITA (364-369 a perfect 6 seed match)

```
STarMir Site 19

    5'->3'        A   UUUUAU     UC     N
    Target   348   UCU       CAUGA  CUACC    368
                   |||       |||||  |||||
    miRNA     22   GGG       GUACU  GAUGG    1
    3'->5'             UUGUU       UU     AU
```

PITA predicts 22 binding sites with an overall score -12.08.
There is one binding site covering the SNP position:

Gene	microRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-196a	364-369	6:0:0	-17.1	-5.04	-12.05

The following site may cause the opening of the SNP position:

Gene	miRNA	Position	Seed	dGduplex	dGopen	ddG
A-Allele	hsa-miR-196a	368-373	6:1:1	-9.4	-3.22	-6.17

FindTar predictions: there is only one binding site predicted for the C-Allele while there are 2 binding sites for the A-Allele; the additional one covers the SNP position:

Position	Structure	Loop Score	ΔG
349-370	3' GGGUUGUUGUACUUUGAUGGAU 5' |::**:|*|||||**||||||| 5' CTTTTTATCATGATCCTACCTA 3'	15.00	-21.50

Appendix C

MSbind data

Contents

1. LIG3 (L = 124)

NM_002311.4 vs. rs4796030
miRNA: miR-221
SNP Pos. 83

	A-allele	C-allele
Target Site:	77 - 98	80 - 98

RNAsubopt and Barrier Setting

Offset	Barrier	A-allele	C-allele	A-allele	C-allele
6.0	1.2	20,646	28,997	349	317

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
A-allele	-20.30	-3.39	-11.77	-8.53
C-allele	-24.60	-0.20	-15.96	-8.64

Local Minima

MFE	A -allele	C-allele
#Pairings	9	7
Opening Energy	-10.48	-12.68
Structure Energy	-33.70	-35.90
MFE Barrier	6.0	6.00
Minima		
Less Pairings	28.4% 99	6.3% 20
Equal Pairings	38.5% 134	10.4% 33
Greater Pairings	33.0% 115	83.2% 263
Identical to MFE	4.0% 14	8.5% 27
Minima (-MFE)	**348**	**316**
Avg. Opening of All Local Minima (excluding MFE)	**-10.38**	**-12.33**

Less/Equal Pairings than MFE

#Pairings	#Minima A-allele	Target Site Approx. Avg.	#Minima C-allele	Target Site Approx. Avg.
4	0.29% 1	-7.40	1.27% 4	-5.99
5	1.15% 4	-10.49	2.22% 7	-6.98
6	4.02% 14	-11.07	2.85% 9	-10.00
7	8.62% 30	-8.36	**10.44% 33**	**-12.41**
8	14.37% 50	-10.01	**24.68% 78**	**-13.88**
9	**38.51% 134**	**-9.79**	**16.77% 53**	**-11.50**
Total	**66.95% 233**	**-9.73**	**58.23% 184**	**-12.31**
Avg. Barrier	**1.78**		**1.72**	

Local Minima with Less Pairings

	A-allele	C-allele
#LM Less Pairings:	28.4% 99	6.33% 20
Avg. Target Site Energy	-9.65	-8.14
Less Pairings Avg. Barrier Height	1.69	1.74

Local Minima with Less or Equal Approx. Target Site Energy than the MFE

	A -allele	C-allele
Less	61.5% 214	57.9% 183
Equal	4.0% 14	4.4% 14
Less or Equal	65.5% 228	62.3% 197

N+ Test

A-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-30.20	15 **(105)**	2*2.50, 2.30, 2.20, 2*2.10, 2*1.80, 5*1.30, 2*1.20	1.75	-21.08, -16.28, -13.47, -11.96, -10.89, -10.56, -10.49, -10.06, -9.98, 2*-9.76, 2*-8.46, -8.21, -6.98	-11.09
-30.30	14	2*2.60, 2.10, 5*1.80, 1.70, 2*1.60, 3*1.30	1.79	-11.53, -10.76, -7.08, -9.98, 2*-9.86, 2*-10.48, 2*-10.59, -12.03, -6.85, -9.18, -22.36	-10.83
-30.40	2	2.20, 1.80	2.00	-10.36, -9.96	-10.16
-30.50	7	2*2.80, 2.70, 2.40, 2.20, 2.10, 1.80	2.40	-10.36, -10.86, -10.46, -8.51, -9.38, -10.06, -8.76	-9.77
-30.60	8 **(67)**	2*2.90, 2.40, 2.20, 1.80, 1.50, 2*1.30	2.04	-11.06, -10.89, -10.56, -10.48, -10.46, -9.18, -5.85, -7.38	-9.48
-30.70	4	3.00, 1.80, 1.30, 1.60	1.93	-10.56, -10.48, -7.48, -9.18	-9.43
-30.80	4	1.50, 1.30, 1.80, 1.60	1.55	-10.46, -7.58, -10.36, -10.48	-9.72
-30.90	8	2*3.20, 2.2, 1.90, 1.80, 2*1.30, 1.40	2.04	3*-7.68, -9.38, -10.86, -11.33, -10.46, -9.78	-9.36
-31.00	3	1.40, 3.30, 1.80	2.17	-7.78, -10.86, -10.56	-9.73
-31.10	5	2*2.30, 1.80, 1.60, 1.30	1.86	-11.53, -9.98, -9.18, 2*-7.88	-9.29
-31.20	4	1.31, 2.30, 3.50, 1.60	2.18	-12.43, -7.98, -14.47, -9.46	-11.09
-31.30	4	2*1.80, 1.30, 2.60	1.43	-10.86, -9.38, -9.78, -8.08	-9.53
-31.40	2	2.60, 1.90	2.25	2*-8.18	-8.18
-31.50	4	2*1.30, 2.10, 2.30	1.43	-10.49, -11.96, -9.98, -9.76	-10.55
-31.60	6 **(21)**	3.70, 3.90, 2*1.30, 2.00, 3.20	2.57	-12.03, -6.85, -10.48, -9.86, 2*-10.59	-10.06
-31.70	2	2.30, 1.30	1.80	-9.96, -9.78	-9.87
-31.80	1	2.30	2.30	-10.06	-10.06

-31.90	2	2.10, 1.80	1.95	-10.89, -9.98	-10.44
-32.00	1 **(10)**	4.30	4.30	-10.48	-10.48
-32.10	1	2.20	2.20	-10.36	-10.36
-32.20	1	2.40	2.40	-10.46	-10.46
-32.30	1	3.20	3.20	-10.56	-10.56
-32.40	2	1.80, 1.60	1.70	-10.48, -9.18	-9.83
-32.60	2	4.70, 3.70	4.20	-10.86, -9.38	-10.12
-33.00	1	1.30	1.30	-9.78	-9.78
-33.20	1	2.30	2.30	-9.98	-9.98
MFE: -33.70	1	6.00	6.00	-10.48	-10.48
-3283.9 Avg. -30.98	**106**	**220.81**	**2.08**	**-1073.4**	**-10.13**

A-allele - N+ Averages

Test #LM-MFE	Avg. Structure Energy	Avg. Barrier	Avg. Opening Energy
10	-324.8/10 -32.48	27.5/10 2.75	-101.52/10 10.15
21	-673.4/21 -32.07	52.7/21 2.51	-212.59/21 -10.12
67	-2098.7/67 -31.32	142.71/67 2.13	-656.52/67 -9.80
105	-3250.2/105 -30.95	214.81/105 2.04	-1063.26/105 -10.13

N+ Test

C-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-32.40	9 **(107)**	2* 2.50, 2*2.40, 2*1.80, 1.60, 2*1.20	1.83	-16.20, 2*-15.30, -14.43, -12.76, -12.36, -10.88, -10.48, -6.30	-12.67
-32.50	7	2*2.60, 2.40, 2.10, 1.80, 1.60, 1.20	2.04	-15.30, -14.53, -12.46, 2*-12.68, -12.38, -12.36	-13.20
-32.60	9	3* 2.70, 2.40, 3*2.10, 2.00, 1.80	2.29	-15.87, -13.29, -12.46, -12.38, -11.59, -11.20, -10.86, -10.61, -6.80	-11.67
-32.70	8	2.80, 2.40, 2.30, 2.10, 2*1.60, 2*1.20	1.90	-20.23, -16.20, -13.16, -12.38, -11.58, -10.96, -9.48, -7.95	-12.37
-32.80	8	2*2.90, 2.30, 1.90, 2*1.80, 1.50, 1.40	2.06	-12.76, -12.68, -12.36, -11.68, -11.20, -10.88, 2*-9.58	-11.34
-32.90	8 **(66)**	2* 3.00, 2.70, 3*1.80, 1.30, 1.20	2.08	-16.23, -13.50, -12.76, -12.68, -12.46, -11.78, -11.16, -9.68	-12.53
-33.00	7	2*3.10, 2.10, 1.80, 1.60, 2* 1.30	2.04	-13.60, -13.29, -12.68, -11.26, 2*-9.78, -8.25	-11.23
-33.10	1	1.60	1.60	-11.58	-11.58
-33.20	5	3.20, 1.80, 1.40,	1.78	-16.50, -15.46, -15.30, -12.76, -11.68	-14.34

		1.30, 1.20			
-33.30	6	3.40, 2.20, 1.90, 1.80, 1.40, 1.20	1.98	-16.23, -13.50, -11.78, 3*-10.08	-11.96
-33.40	4	2.10, 1.80, 1.30, 1.20	1.60	-18.16, -17.13, -16.23, -13.60	-16.28
-33.50	4	3.60, 3.20, 1.80, 1.60	2.55	-13.93, -12.38, -11.58, -10.28	-12.04
-33.60	3	3.70, 1.60, 1.40	2.23	-15.30, -11.86, -11.68	-12.95
-33.70	6	3.80, 2.60, 2*1.80, 2*1.20	1.56	-16.23, -15.30, -14.43, -13.50, -11.78, -10.48	-13.62
-33.80	3 **(22)**	3.20, 1.80, 1.20	2.07	-14.53, -13.60, -12.68	-13.60
-33.90	1	4.00	4.00	-12.38	-12.38
-34.00	2	1.80, 1.20	1.50	-16.20, -15.30	-15.75
-34.10	2	2*2.40	2.40	-12.36, -10.88	-11.62
-34.20	2	4.30, 3.20	3.75	-12.68, -12.46	-12.57
-34.30	2	2.10, 1.80	1.95	-13.29, -12.38	-12.84
-34.50	1 **(10)**	4.60	4.60	-12.76	-12.76
-34.60	1	1.80	1.80	-12.68	-12.68
-34.80	1	1.60	1.60	-11.58	-11.58
-34.90	1	1.40	1.40	-11.68	-11.68
-35.00	3	4.20, 1.80, 1.20	2.40	-16.23, -13.50, -11.78	-13.84
-35.10	1	2.10	2.10	-13.60	-13.60
-35.30	1	5.40	5.40	-15.30	-15.30
-35.60	1	5.20	5.20	-12.38	-12.38
MFE: -35.90	1	6.00	6.00	-12.68	-12.68
-3594.4 Avg. -33.28	**108**	**237.7**	**2.20**	**1374.8**	**-12.73**

C-allele - N+ Averages

Test #LM-MFE	**Avg. Structure Energy**	**Avg. Barrier**	**Avg. Opening Energy**
10	-349.8/10 -34.98	29.3/10 2.93	131.49/10 -13.15
22	-758.3/22 -34.47	58.7/22 2.67	-290.23/22 -13.19
66	-2222/66 -33.67	147.7/66 2.24	-858.00/66 -13.00
107	-3558.5/107 -33.26	231.7/107 2.17	-1362.12/107 -12.73

Ordered by Deepest Local Minima

A-allele

Barrier	#LM	Opening	Avg. Opening
4.70	1	-10.86	-10.86
4.30	1	-10.48	-10.48
3.90	1	-6.85	-6.85
3.70	2	-12.03, -9.38	-10.71
3.50	1	-14.47	-14.47
3.30	1	-10.86	-10.86
3.20	4 **(11)**	-9.38, -10.48, -10.56, -10.86	-10.32
3.00	1	-10.56	-10.56
2.90	2	-10.56, -10.48	-10.52
2.80	2	-9.38, -10.86	-10.12
2.70	1	-8.76	-8.76
2.60	4 **(21)**	-22.36, -8.08, -8.18, -11.53	-12.54
2.50	2	-21.08, -10.56	-15.82
2.40	5	4*-10.46, -5.35	-9.44
2.30	14	-7.98, -8.26, 3*-9.96, 4*-9.98, 3*-10.06, -11.53, -20.88	-10.62
2.20	9	-5.85, -6.85, -7.68, 4*-10.36, -12.03, -13.47	-9.70
2.10	16 **(67)**	-5.05, -6.68, -8.21, -8.51, -9.98, -9.09, 3*-10.89, -10.48, 3*-11.96, -15.31, -15.61, -22.36	-11.24
2.00	7	-8.18, -8.26, -10.06, 2*-10.59, -16.66, -21.08	-10.69
1.90	8	-7.68, -8.08, 2*-8.18, -9.96, -10.86, -11.69, -15.98	-10.08
1.80	42 **(124)**	2*-6.85, 2*-7.68, -7.76, 2*-7.88, 2*-7.98, -8.08, 2*-8.18, 3*-9.38, -9.96, 3*-9.98, -10.06, -10.36, -10.46, 3*-10.48, -10.56, 3*-10.59, -10.76, 4*-10.86, -10.89, -11.53, -11.96, 2*-12.03, 2*-14.47, -20.88	-10.19

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
11	39.9/11 3.63	-116.21/11 -10.56
21	67.4/21 3.21	-226.96/21 -10.81
67	170.0/67 2.54	-721.57/67 -10.77
124	274.8/124 2.22	-1315.73/124 -10.61

C-allele

Barrier	#LM	Opening	Avg. Opening
5.40	1	-15.30	-15.30
5.20	1	-12.38	-12.38
4.60	1	-12.76	-12.76
4.30	1	-12.68	-12.68
4.20	1	-11.78	-11.78
4.00	1	-12.38	-12.38
3.80	1	-15.30	-15.30
3.70	1	-15.30	-15.30
3.60	1	-13.93	-13.93
3.40	1 **(10)**	-11.78	-11.78
3.20	4	-12.38, -12.46, -12.68, -15.30	-13.21
3.10	2	-8.25, -11.26	-9.76
3.00	2	-11.78, -12.76	-12.27
2.90	1	-12.68	-12.68
2.80	1 **(20)**	-20.23	-20.23
2.70	4	-6.80, -11.16, -12.46, -15.87	-11.57
2.60	3	-10.48, -12.38, -12.46	-11.77
2.50	2	-6.30, -12.76	-9.53
2.40	10	-6.86, -9.08, -10.86, 2*-10.88, -10.96, 3*-12.36, -15.30	-11.19
2.30	4	-9.48, -11.20, -15.30, -20.23	-14.05
2.20	4	-10.08, -10.73, -12.46, -15.13	-12.10
2.10	18 **(65)**	-7.20, -10.48, -10.61, -10.88, -11.59, -12.36, -12.38, -12.43, -12.68, -13.16, 3*-13.29, 4*-13.60, -17.71	-12.54
2.00	4	-10.16, -11.20, -11.78, -15.30	-12.11
1.90	10	-6.00, 4*-8.58, -8.60, -9.33, -9.58, -10.08, -13.93	-9.18
1.80	43	-6.95, -8.25, -8.88, -9.96, 2*-10.08, 3*-10.28, -10.48, -10.88, -11.26, 2*-11.78, -12.36, 3*-12.38, -12.46, 3*-12.68, -12.76, -13.29, 4*-13.60, -13.93, 4*-15.30, 9*-13.23, -20.23	-12.47

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	42.2/10 4.22	-133.59/10 -13.36
20	73/20 3.65	-263.37/20 -13.17
65	177.1/65 2.72	-811.86/65 -12.49
122	281.8/122 2.31	-1509.76/122 -12.38

Ordered by Opening Energy

A-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-4.15	1	1.20	1.20
-4.35	1	1.30	1.30
-4.55	1	1.60	1.60
-5.05	1	2.10	2.10
-5.35	2	2.40, 1.40	1.90
-5.68	1	1.20	1.20
-5.85	3 **(10)**	2.20, 1.60, 1.20	1.67
-5.98	3	1.30, 1.40, 1.50	1.40
-6.18	1	1.70	1.70
-6.38	1	1.20	1.20
-6.48	3	1.50, 1.30, 1.20	1.33
-6.58	1	1.70	1.70
-6.68	2 **(21)**	2.10, 1.50	1.80
-6.85	4	3.90, 2.20, 2*1.80	2.43
-6.88	1	1.50	1.50
-6.98	1	1.20	1.20
-7.01	1	1.30	1.30
-7.08	3	1.70, 1.30, 1.20	1.40
-7.16	1	1.20	1.20
-7.26	1	1.30	1.30
-7.38	3	2*1.30, 1.20	1.27
-7.40	1	1.30	1.30
-7.46	2	1.30	1.30
-7.48	3	3*1.30	1.30
-7.56	3	1.60, 1.50, 1.30	1.47
-7.58	3	3*1.30	1.30
-7.66	1	1.70	1.70
-7.68	10	2.20, 1.90, 2*1.80, 2*1.50, 3*1.40, 1.30	1.62
-7.76	1 **(60)**	1.80	1.80
-7.78	4	3*1.40, 1.20	1.35
-7.86	2	1.30, 1.20	1.25
-7.88	8	2*1.80, 1.70, 5*1.30	1.48
-7.91	1	1.60	1.60
-7.96	1	1.30	1.30
-7.98	4	2.30, 2*1.80, 1.40	1.83
-8.06	1	1.30	1.30
-8.08	4	2.60, 1.90, 1.80, 1.50	1.95
-8.16	1	1.40	1.40
-8.18	8	2.60, 2.00, 2*1.90, 2*1.80, 2*1.60	1.90
-8.21	2	2.10, 1.20	1.65
-8.26	2	2.30, 2.00	2.15
-8.33	2 **(100)**	1.40, 1.30	1.35

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	16.2/10 1.62	-52.03/10 -5.20
21	32.6/21 1.55	-121.91/21 -5.81
60	92.6/60 1.54	-410.38/60 -6.84
100	158.5/100 1.59	-731.63/100 -7.32

C-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-5.50	1	1.70	1.70
-5.66	1	1.20	1.20
-6.00	1	1.90	1.90
-6.06	1	1.60	1.60
-6.30	2	2.50, 1.20	1.85
-6.35	2	2*1.20	1.20
-6.40	1	1.30	1.30
-6.45	2 **(11)**	2*1.30	1.30
-6.80	2	2.70, 1.40	2.05
-6.86	1	2.40	2.40
-6.95	1	1.80	1.80
-7.20	1	2.10	2.10
-7.88	1	1.20	1.20
-7.95	1	1.20	1.20
-7.98	1	1.30	1.30
-8.08	1 **(20)**	1.30	1.30
-8.25	3	3.10, 1.80, 1.40	2.10
-8.28	1	1.30	1.30
-8.58	4	4*1.90	1.90
-8.60	1	1.90	1.90
-8.68	1	1.50	1.50
-8.88	1	1.80	1.80
-8.93	1	1.40	1.40
-8.98	1	1.70	1.70
-9.08	1	2.40	2.40
-9.30	1	1.70	1.70
-9.33	1	1.90	1.90
-9.48	2	2.30, 1.50	1.90
-9.56	3	3*1.40	1.40
-9.58	6	1.90, 1.60, 2*1.50, 2*1.20	2.23
-9.60	1	1.20	1.20
-9.66	1	1.50	1.50
-9.68	3	3*1.30	1.30

-9.76	2	1.60, 1.30	1.45
-9.78	6 **(60)**	6*1.30	1.30
-9.86	2	2*1.30	1.30
-9.96	2	1.80, 1.40	1.60
-10.08	12	2.20, 1.90, 2*1.80, 2*1.70, 3*1.40, 3*1.30	1.60
-10.13	1	1.20	1.20
-10.16	1	2.00	2.00
-10.23	1	1.30	1.30
-10.28	4	3*1.80, 1.50	1.73
-10.48	4	2.60, 2.10, 1.80, 1.70	2.05
-10.61	2	2.10, 1.40	1.75
-10.73	1	2.20	2.20
-10.86	2	2.40, 1.40	1.90
-10.88	6	2*2.40, 2.10, 1.80, 1.30, 1.20	1.87
-10.96	3 **(101)**	2.40, 1.50, 1.20	1.70

C-allele - Averages

Test #LM-MFE	**Avg. Barrier**	**Avg. Opening Energy**
11	16.4/11 1.49	-67.82/11 -6.17
20	31.8/20 1.59	-134.32/20 -6.72
60	95.5/60 1.59	-505.07/60 -8.42
101	165.9/101 1.64	-931.06/101 -9.22

2. CBR1 (L = 284)

NM_001757.2 vs. rs9024
miRNA: miR-574-5p
SNP Pos. 133
Target site: 121 - 162

RNAsubopt and Barrier Setting

Offset	Barrier	G-allele	A-allele	G-allele	A-allele
6.0	1.4	10,987,436	16,209,366	7,457	11,187

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_disrupt	dG_total
G-allele	-27.80	-5.36	-8.09	-19.72
A-allele	-27.80	-5.22	-5.96	-21.84

Local Minima

MFE	G-allele	A-allele
#Pairings	14	12
Opening Energy	-10.90	-9.90
Structure Energy	-52.10	-51.10
MFE Barrier	6.00	6.00
Minima		
Less Pairings	39.6% 2,954	20.3% 2,274
Equal Pairings	18.6% 1,385	30.7% 3,431
Greater Pairings	41.8% 3,117	49.0% 5,481
Identical to MFE	12.8% 951	8.5% 951
Minima (- MFE)	**7,456**	**11,186**
Approx. Target Energy Avg. (excluding MFE)	**-10.58**	**-9.27**

Less/Equal Pairings than MFE Target Site

#Pairings	#Minima G-Allele	Target Site Approx. Avg.	#Minima A-allele	Target Site Approx. Avg.
4	-	-	0.07% 8	-2.20
6	0.28% 21	-4.39	0.61% 68	-4.62
7	0.25% 19	-4.57	1.25% 140	-4.48
8	0.15% 11	-5.86	3.22% 360	-5.68
9	0.52% 39	-5.23	0.75% 84	-5.81
10	1.44% 107	-5.64	8.31% 929	-7.10
11	1.22% 91	-6.48	6.12% 685	-7.36
12	24.75% 1,845	-7.64	**30.67% 3,431**	**-8.25**
13	11.01% 821	-8.38	**14.44% 1,615**	**-8.55**
14	**18.58% 1,385**	**-10.68**	**3.41% 382**	**-9.00**
Total Minima	**58.19% 4,339**	**-8.62**	**68.85% 7,702**	**-7.88**
Avg. Barrier	**1.83**		**1.82**	

Local Minima with Less Pairings

	G-allele	A-allele
#LM Less Pairings:	36.6% 2,954	20.3% 2,274
Avg. Target Site Energy	-7.66	-6.65
Less Pairings Avg. Barrier Height	1.82	1.82

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	G-allele	A-allele
Less	56.37% 4,203	58.04% 6,492
Equal	13.37% 997	8.93% 999
Less or Equal	69.74% 5,200	66.97% 7,491

N+ Test

G-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-49.80	25 **(123)**	10*3.70, 2*3.20, 2.10, 2.30, 2*2.40, 7*2.20, 2*1.70	2.86	5*-18.70, 5*-12.75, -11.10, 6*-10.90, 5*-8.20, 2*-7.30, -7.00	-11.85
-49.90	23	2*3.80, 3.70, 3.40, 2*3.20, 2.90, 4*2.20, 2*2.40, 3*2.10, 2.00, 2*1.90, 1.80, 2*1.70, 1.60	2.46	5*-12.75, -12.00, -11.10, 5*-10.90, -9.20, 5*-8.20, 2*7.30, 3*-7.00	-9.88
-50.00	15	3*3.90, 3.80, 3.40, 3.20, 2.40, 2*2.30, 2.10, 2*2.20, 3*1.80	3.77	2*-18.70, 2*-12.75, 4*-10.90, 9.80, 2*-8.20, 2*-7.60, -7.30, -7.00	-10.81
-50.10	13 **(60)**	5*4.00, 3.90, 3.80, 2*3.20, 3*2.10, 1.90	3.25	2*-12.75, -11.60, -11.10, 3*-8.20, -7.30, 3*-7.00, 2*-5.80	-8.67
-50.20	9	2*4.10, 3.70, 2*2.90, 3*2.40, 1.80	2.97	2*-18.70, -12.75, -10.90, -9.00, -8.20, 3*-7.30	-11.13
-50.30	6	4.20, 3.90, 3.80, 3.70, 2.90, 2.10	3.43	2*-18.70, -12.75, -10.90, -8.20, -7.30	-12.76
-50.40	10	4*4.30, 2*3.80, 3.40, 2*2.90, 2.10	3.61	3*-12.75, -9.20, 3*-8.20, 3*-7.30	-9.40
-50.50	6 **(22)**	2*4.40, 3.80, 2.90, 2.30, 1.80	3.27	2*-18.70, -12.75, -10.90, -8.20, -7.60	-12.81
-50.60	8 **(16)**	4*4.50, 2*3.80, 2*2.10	3.73	3*-12.75, 3*-8.20, -10.90, -5.80	-9.94
-50.70	1	1.90	1.90	-10.90	-10.90
-50.80	4	4.70, 3.90, 2.30,	3.25	4*-10.90	-10.90

		2.10			
-51.30	2	3.90, 2.30	3.10	2*-10.90	-10.90
-51.60	1	5.50	5.50	-10.90	-10.90
MFE: -52.10	1	6.00	6.00	-10.90	-10.90
-6219.6 Avg. -50.16	**124**	**376.6**	**3.04**	**1323.55**	**-10.67**

G-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
16	-812.9/18 -50.80	56.4/16 3.53	166.75/16 -10.42
60	-3024.8/60 -50.41	201.4/60 3.36	-626.95/60 -10.45
22	-1115.9/22 -50.72	76/22 3.45	-243.6/22 -11.07
123	-6167.5/123 -50.14	370.6/123 3.01	-1312.65/123 -10.67

A-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-49.20	20 **(113)**	4.10, 3.70, 2*2.90, 3*2.40, 2.20, 3*2.00, 5*1.90, 4*1.80	2.29	2*-17.70, -11.75, 3*-10.42, -9.90, -8.00, 2*-7.60, 3*-7.30, -7.20, 4*-6.30, 2*-5.80	-8.87
-49.30	19	2*4.20, 3.90, 2*3.80, 3.70, 2.90, 2*2.80, 2*2.20, 2*2.10, 2*2.00, 1.90, 3*1.80	2.74	2*-17.70, -11.75, -10.42, -9.90, 2*7.60, 7*-7.30, -7.20, -6.30, 2*-5.80, -4.50	-8.60
-49.40	16 **(74)**	3*4.30, 2*3.80, 3.40, 2*2.90, 2.20, 3*2.00, 3*1.90, 1.80	2.84	3*-11.75, 3*-10.42, -8.20, -7.60, 3*7.20, 3*6.30, 2*5.80	-8.40
-49.50	11	3*4.40, 3.80, 2*3.10, 2.90, 2.30, 2.20, 2.10, 1.80	3.15	2*-17.70, -11.75, -10.55, -9.90, -7.60, 3*-7.30, -7.20, -5.80	-10.00
-49.60	14	3*4.50, 4.30, 3.90, 2*3.80, 3.40, 3*2.20, 2*2.10, 1.80	3.24	3*11.75, 9.90, -7.60, 5*-7.30, 3*7.20, -5.80	-8.33
-49.70	7	2.90, 2.40, 2*2.20, 2.10, 1.90, 1.70	2.20	-9.90, 5*-7.30, -5.80	-7.46
-49.80	6	4.70, 2*3.90, 2.30, 2.20, 2.10	3.18	4*-9.90, 2*-7.30	-9.03
-49.90	2 **(20)**	4.80, 2.10	3.45	2*-7.30	-7.30

-50.00	3	2.40, 2*1.80	2.00	2*-7.60, -7.30	-7.50
-50.10	3	2*5.00, 3.90	4.63	-7.30, 2*-5.80	-6.30
-50.20	3 **(12)**	2*5.10, 2.90	4.36	3*-7.30	-7.30
-50.30	3	5.20, 3.90, 2.30	3.80	2*-9.90, -7.30	-9.03
-50.40	3	2*5.30, 2.10	4.23	3*-7.30	-7.30
-50.50	1	1.80	1.80	-7.60	-7.60
-50.60	2	2*5.50	5.50	-9.90, -5.80	-7.85
MFE: -51.10	1	6.00	6.00	-9.90	-9.90
-5652.3 Avg. -49.58	**114**	**340.2**	**2.98**	**-968.44**	**-8.57**

A-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
12	-604.4/12 -50.37	50/12 4.17	94.2/12 -7.85
20	-1004.5/20 -50.23	76.8/20 3.84	-150.2/20 -7.51
74	-3680.5/74 -49.74	236.5/74 3.20	-617.76/74 -8.35
113	-5601.2/113 -49.57	334.2/113 2.96	-958.54/113 -8.48

Ordered by Deepest Local Minima

G-allele

Barrier	#LM	Opening	Avg. Opening
5.50	1	-10.90	-10.90
4.70	1	-10.90	-10.90
4.50	4	-5.80, -10.90, 2*-12.75,	-10.55
4.40	2	-18.70, -12.75	-15.73
4.30	4 **(12)**	2*-7.30, 2*-12.75	-10.02
4.20	1	-7.30	-7.30
4.10	2	2*-7.30	-7.30
4.00	5 **(20)**	2*-5.80, -11.10, -11.60, -12.75	-9.41
3.90	7	-7.30, 3*-10.90, 2*-12.75, -18.70	-12.03
3.80	10	-7.30, 8*-8.20, -10.90	-8.38
3.70	13	-7.30, 2*-8.20, 3*-10.90, 2*-11.10, 2*-12.75, 3*-18.70	-12.32
3.60	4	4*-10.90	-10.90
3.50	12 **(66)**	2*-7.30, 7*-10.90, -11.50, -11.60, -18.70	-11.06
3.40	14	-5.80, -8.20, 3*-9.20, 2*-9.80, -10.00, -10.55, 2*-10.90, -12.75, 2*-18.70	-10.98
3.30	12	-5.80, -7.90, -8.20, 2*-9.93, -10.20, 2*-10.90, -11.10, -11.30, -12.75, -18.70	-10.63
3.20	36 **(128)**	-4.50, -5.80, 10*-7.00, -7.30, 5*-8.20, -9.10, -9.93, -10.00, 2*-10.90, -11.60, 5*-12.75, 7*-18.70	-10.71

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
12	54.2/12 4.52	-135.55/12 -11.30
20	86.6/20 4.33	-204.5/20 -10.23
66	256.4/66 3.88	-709/66 -10.74
128	458.80/128 3.58	-1375.99/128 -10.75

A-allele

Barrier	#LM	Opening	Avg. Opening
5.50	2	-5.80, -9.90	-7.85
5.30	2	2*-7.30	-7.30
5.20	1	-7.30	-7.30
5.10	2	2*-7.30	-7.30
5.00	2	2*5.80	-5.80
4.80	1 **(10)**	-7.30	-7.30
4.70	1	-9.90	-9.90
4.50	3	-9.90, 2*-11.75	-11.13
4.40	3	-10.55, -11.75, -17.70	-13.33
4.30	4 **(21)**	-5.80, -7.30, 2*-11.75	-9.15
4.20	2	-5.80, -7.30	-6.55
4.10	1	-7.30	-7.30
4.00	8	-5.80, 4*-7.30, -10.10, -10.60, -11.75	-8.43
3.90	16	7*-7.30, -8.40, 3*-9.90, 2*-10.55, 2*-11.75, -17.70	-9.47
3.80	15 **(63)**	-4.50, 2*-5.80, 8*-7.20, 3*-7.30, -9.90	-7.03
3.70	19	2*-4.50, 2*-7.20, 4*-7.30, -8.40, 3*-9.90, 2*-10.10, 2*-11.75, 3*-17.70	-9.87
3.60	11	-4.90, -5.80, 4*-7.30, -8.40, 4*-9.90	-7.99
3.50	13 **(106)**	-4.50, 2*-7.30, 7*-9.90, -10.50, -10.60, -17.70	-9.78

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	51.8/10 5.18	-71.1/10 -7.11
21	100.4/21 4.78	-191/21 -9.10
63	264.3/63 4.20	-535.85/63 -8.51
106	419.7/106 3.96	-938.45/106 -8.85

Ordered by Opening Energy

G-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-2.20	1	1.60	1.60
-2.90	1	1.60	1.60
-3.40	4	3*1.50, 1.40	1.48
-3.50	4 **(10)**	1.90, 2*1.70, 1.50	1.70
-3.90	10 **(20)**	4*1.70, 3*1.60, 1.50, 2*1.40	1.59
-4.00	9	2*1.70, 1.50, 6*1.40	1.48
-4.10	12	2.80, 2*2.30, 1.90, 1.80, 1.70, 2*1.60, 2*1.50, 2*1.40	1.82
-4.50	26 **(67)**	3.20, 2*2.70, 2.50, 2.30, 2.20, 2.10, 4*2.00, 1.90, 3*1.80, 2*1.70, 2*1.60, 3*1.50, 4*1.40	1.91
-4.70	6	2.40, 2*1.90, 2*1.50, 1.40	1.77
-4.90	8	2.60, 2*2.10, 1.70, 1.60, 1.50, 2*1.40	1.80
-5.00	10	2.70, 2*2.20, 1.80, 1.70, 1.60, 2*1.50, 2*1.40	1.80
-5.20	3	3*1.50	1.50
-5.30	18 **(112)**	1.80, 8*1.60, 2*1.50, 7*1.40	1.52

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	15.9/10 1.59	-32.7/10 -3.27
20	31.8/20 1.59	-71.7/20 -3.59
67	116.6/67 1.74	-273.9/67 -4.09
112	191.5/112 1.71	-502.3/112 -4.48

A-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-2.20	8	5*1.60, 1.50, 2*1.40	1.36
-2.40	1	1.40	1.40
-2.60	1 **(10)**	1.60	1.60
-2.90	17 **(27)**	3*1.90, 5*1.70, 3*1.60, 2*1.50, 4*1.40	1.62
-3.00	6	6*1.40	1.40
-3.40	8	2.40, 2*1.90, 1.50, 4*1.40	1.66
-3.50	26 **(67)**	4*1.40, 3*1.50, 2*1.60, 2*1.70, 3*1.80, 12*1.90	1.73
-3.60	1	1.40	1.40
-3.70	1	1.50	1.50

-3.90	16	2.60, 2*2.10, 3*1.90, 2*1.70, 2*1.60, 2*1.50, 4*1.40	1.73
-4.00	26 **(111)**	5*1.40, 5*1.50, 6*1.60, 2*1.70, 2*1.80, 3*1.90, 2*2.20, 2.70	1.69

A-allele - Averages

Test #LM-MFE	**Avg. Barrier**	**Avg. Opening Energy**
10	15.3/10 1.53	-22.6/10 -2.26
27	42.9/27 1.59	-71.9/27 -2.66
67	109.5/67 1.63	-208.1/67 -3.11
111	184/111 1.66	-381.8/111 -3.44

3. HTR3E (L = 302)

NM_001256614.1 vs. rs56109847
miRNA: miR-510-5p
SNP Pos. 76

	G-allele	A-allele
Target Site:	50 - 80	50 - 80

RNAsubopt and Barrier Setting

Offset	Barrier	G-allele	A-allele	G-allele	A-allele
6.0	1.4	260,869	239,418	1,174	997

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_disrupt	dG_total
G-allele	-28.70	-5.505	-12.645	-16.055
A-allele	-22.50	-2.464	-8.261	-14.239

Local Minima

MFE	G -allele		A-allele	
#Pairings	15		9	
Opening Energy	-24.47		-10.96	
Structure Energy	-66.90		-63.90	
MFE Barrier	6.00		6.00	
Minima				
Less Pairings	10.91%	128	7.53%	75
Equal Pairings	78.86%	925	24.29%	242
Greater Pairings	10.23%	120	68.17%	679
Target Site Identical to MFE	13.73%	161	13.05%	130
Minima (- MFE)	**1,173**		**996**	
Avg. Opening of All Local Minima (excluding MFE)	**-21.74**		**-10.56**	

Less/Equal Pairings than MFE

#Pairings	#Minima G-allele		Target Site Approx. Avg.	#Minima A-allele		Target Site Approx. Avg.
6	0.09%	1	-5.40	0.10%	1	-2.20
7	-		-	0.50%	5	-5.04
8	1.02%	12	-9.42	6.93%	69	-6.03
9	0.34%	4	-10.58	24.3%	242	-9.43
10	0.68%	8	-7.90	**22.5%**	**224**	**-6.67**
11	1.71%	20	-10.50	**4.82%**	**48**	**-7.44**
12	2.30%	27	-9.68	**7.13%**	**71**	**-8.73**
13	1.36%	16	-17.29	**9.84%**	**98**	**-10.88**
14	3.41%	40	-21.63	**20.38%**	**203**	**-18.10**
Total:	**10.91%**	**128**	**-14.35**	**96.5%**	**961**	**-10.34**
Avg. Barrier:	**1.79**			**1.79**		

Local Minima with Less Pairings

	G-allele	A-allele
#LM Less Pairings:	10.91% 128	7.53% 75
Avg. Target Site Energy	-14.35	-5.91

Local Minima with Less or Equal Approx. Target Site Energy than the MFE

	G-allele	A-allele
Less	36.15% 424	595
Equal	27.54% 323	130
Less or Equal	63.68% 747	725

N+ Test

G-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-64.40	12 **(108)**	4*2.10, 4*1.70, 4*1.50	1.77	4*-25.40, 8*-21.37	-22.71
-64.50	14	4*2.10, 2*1.90, 4*1.60, 4*1.50	1.76	8*-21.47, 6*-21.37	-21.43
-64.60	10	4*2.10, 4*1.60, 2*1.90	1.86	6*-21.47, 4*-21.37	-21.43
-64.70	4	4*2.10	2.10	4*-21.47	-21.47
-64.80	4	4*2.30	2.30	4*-21.37	-21.37
-64.90	10 **(64)**	4*2.90, 6*2.30	2.54	4*-21.47, 6*-21.37	-21.41
-65.00	6	6*2.60	2.60	6*-21.47	-21.47
-65.10	8	4*2.30, 4*1.70	2.00	4*-25.30, 4*-21.37	-23.44
-65.20	8	4*2.60, 4*1.70	2.15	4*-25.40, 4*-21.47	-23.44
-65.30	4	4*2.30	2.30	4*-21.37	-21.37
-65.40	4	4*2.50	2.50	4*-21.47	-21.47
-65.80	4	4*1.70	1.70	4*-25.30	-25.30
-65.90	4 **(20)**	4*1.70	1.70	4*-25.40	-25.40
-66.10	4	4*2.10	2.10	4*-21.37	-21.37
-66.20	4 **(12)**	4*2.10	2.10	4*-21.47	-21.47
-66.80	4	4*2.30	2.30	4*-21.37	-21.37
MFE: -66.90	4	6.00, 4.70, 2*2.90	4.13	4*-21.47	-21.47
Total: -7044.8 Avg. -65.23	**108**	**1.99**		**-22.15**	

G-allele - N+ Averages

Test #LM-MFE	Avg. Structure Energy	Avg. Barrier	Avg. Opening Energy
12	-799.6/12 -66.63	34.1/12 2.84	-257.24/12 -21.44
20	-1327.6/20 -66.38	49.3/20 2.47	-444.32/20 -22.22
64	-4205/64 -65.70	149.5/64 2.34	-1433.96/64 -22.41
108	-7044.8/108 -65.23	231.5/108 2.14	-2392.16/108 -22.15

N+ Test

A-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-61.00	19 **(118)**	2*3.10, 6*2.60, 2.50, 2.30, 2*2.20, 2*2.00, 1.60, 4*1.40	2.22	6*-17.47, 3*-10.96, 2*-10.00, 2*-8.06, -7.90, 3*-7.20, -6.20, -5.80	-11.33
-61.10	9	2.10, 2*1.90, 2*1.80, 2*1.70, 2*1.60	1.79	2*-13.40, 3*-10.96, -8.86, 3*-7.20	-10.02
-61.20	22	5*3.30, 2.80, 4*2.60, 2*2.10, 2*1.80, 4*1.70, 4*1.40	2.27	4*-21.40, 4*-17.47, 4*-10.96, 2*-7.90, 4*-7.20, 2*-6.90, -6.20, -5.50	-12.25
-61.30	6	2*3.40, 2.80, 2.60 2*2.10	2.73	-10.96, 2*-10.40, -9.00, -7.50, -7.20	-9.24
-61.40	10 **(62)**	4*2.50, 3*2.10, 1.50, 2*1.40	2.06	4*-17.47, 2*-10.40, -8.30, -6.20, 2*-5.40	-11.60
-61.50	6	2*3.60, 2*2.10, 2*1.50	2.40	3*-10.96, 3*-7.20	-9.08
-61.60	5	3.70, 2.70, 2.10, 2*1.60	2.34	-10.96, -8.66, 2*-8.30, -7.20	-8.68
-61.70	6	3*2.10, 3*1.40	1.75	-10.96, 2*-10.00, -8.10, -7.90, -7.20	-9.03
-61.80	2	2.30, 2.10	2.20	-8.86, -5.80	-7.33
-61.90	6	4*1.70, 2*2.90	2.10	4*-21.40, -10.96, -7.20	-17.29
-62.00	3	2*2.60, 1.40	2.20	-10.96, 2*-7.70	-8.79
-62.10	4	2*4.20, 2.50, 1.50	3.10	2*-10.40, -8.30, -5.40	-8.63
-62.20	7 **(20)**	2*2.60, 4*2.10, 1.40	2.14	4*-17.47, -10.96, -7.20, -6.20	-13.46
-62.40	3	4.50, 2*2.50	3.17	-10.96, -8.10, -7.20	-8.75
-62.50	1 **(10)**	4.60	4.60	-5.80	-5.80
-62.90	5	5.00, 4.70, 2*2.90, 1.40	3.38	4*-17.47, -6.20	-15.22
-63.20	2	2*2.10	2.10	-10.96, -7.20	-9.08
MFE: -63.90	2	6.00, 5.70	5.85	-10.96, -7.20	-9.08

-6188.5 Avg.	**118**	**277.5**	**2.35**	**1284.53**	**-10.89**

A-allele - N+ Averages

Test #LM-MFE	Avg. Structure Energy	Avg. Barrier	Avg. Opening Energy
10	-631.2/10 -63.12	37.4/10 3.74	-118.2/10 -11.82
20	-1253.8/20 -62.69	61.9/20 3.10	-238.7/20 -11.94
62	-3844.4/62 -62.01	155.1/62 2.50	-682.99/62 -11.02
118	-7267.5/118 -61.59	277.5/118 2.35	-1284.53/118 -10.89

Ordered by Deepest Local Minima

G-Allele

Barrier	#LM	Opening	Avg. Opening
4.70	1	-21.47	-21.47
3.30	2	2*-21.47	-21.47
3.20	2	-23.35, -11.16	-17.26
3.10	3	2*-21.47, -17.70	-20.21
3.00	4 **(12)**	2*-21.47, -10.96, -7.20	-15.27
2.90	12 **(24)**	-23.35, 10*-21.47, -17.70	-21.31
2.80	7	4*-23.35, 2*-21.47, -10.83	-21.02
2.70	6	2*-22.5, 4*-21.47	-21.81
2.60	14	10*-21.47, 4*-10.20	-18.25
2.50	32 **(83)**	4*-23.55, 26*-21.47, 2*-20.21	-21.65
2.40	5	2*-21.47, 2*-19.71, -10.83	-18.64
2.30	82 **(170)**	8*-23.45, 2*-23.25, 2*-22.75, 62*-21.37, 2*-20.11, 2*-20.01, 2*-19.61, 2*-10.83	-21.29

G-Allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
12	39/12 3.25	-220.66/12 -18.39
24	73.8/24 3.08	-476.41/24 -19.85
83	226/83 2.73	-1704.3/83 -20.53
170	426.60/83 2.51	-3543.15/170 -20.84

A-Allele

Barrier	#LM	Opening	Avg. Opening
5.70	1	-7.20	-7.20
5.00	1	-17.47	-17.47
4.70	1	-17.47	-17.47
4.60	1	-5.80	-5.80
4.50	1	-8.10	-8.10
4.20	2	2*-10.40	-10.40
3.70	1	-8.30	-8.30
3.60	2 **(10)**	-10.96, -7.20	-9.08
3.40	2	2*-10.40	-10.40
3.30	5	2*-10.96, 2*-7.20, -6.90	-8.64
3.10	2	-7.20, -10.96	-9.08
3.00	4 **(23)**	-8.10, -8.03, -7.60, -7.10	-7.71
2.90	8	6*-17.47, -10.96, -7.20	-15.37
2.80	5	-9.00, 2*-7.50, -6.90, -6.83	-7.60
2.70	6	-10.96, 2*-8.30, -7.60, -7.50, -7.20	-8.31
2.60	25 **(67)**	10*-17.47, -11.90, 4*-10.96, -9.00, 2*-8.10, -7.63, 2*-7.20, 3*-5.80, -3.80	-11.95
2.50	37 **(104)**	4*-17.47, 7*-10.96, 4*-10.40, 3*-8.30, 2*-8.10, 2*-8.03, 7*-7.20, 4*-7.10, -6.60, -5.50, 2*-5.80	-9.40

A-Allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	43.8/10 4.38	-103.3/10 -10.33
23	85.3/23 3.71	-216.31/23 -9.40
67	203.7/67 3.04	-725.75/67 -10.83
104	296.2/104 2.85	-1074.01/104 -10.33

Ordered by Opening Energy

G-Allele

Opening	#LM	Barrier	Avg. Barrier
-5.40	1	1.70	1.70
-5.50	2	2.10, 1.40	1.75
-5.80	2	2*1.60	1.60
-6.20	2	2.00, 1.40	1.70
-7.20	3 **(10)**	3.00, 2.10, 1.50	1.87
-7.60	1	1.40	1.40
-8.10	2	2*1.50	1.50
-8.50	2	1.90, 1.40	1.65
-9.10	2	2*1.50	1.50
-9.20	6 **(23)**	2*1.60, 4*1.40	1.47
-9.40	2	2*1.80	1.80
-9.43	1	1.40	1.40
-9.60	4	4*1.40	1.40
-9.70	8	4*2.10, 4*1.40	1.75
-10.13	5	2.10, 1.70, 2*1.60 ,1.40	1.68
-10.20	8	4*2.60, 4*1.90	2.25
-10.80	3	1.70, 2*1.60	1.63
-10.83	8 **(62)**	2.80, 2.40, 2*2.30, 2.10, 1.70, 2*1.60	2.10
-10.96	3	3.00, 2.10, 1.50	2.20
-11.16	4	3.20, 2.10, 1.70, 1.50	2.13
-11.70	3	2.20, 1.70, 1.50	1.80
-16.30	2	2*1.70	1.70
-16.70	2	2*1.40	1.40
-17.27	2	2*1.80	1.80
-17.70	8	3.10, 2.90, 2*2.10, 2*1.60, 2*1.40	2.03
-18.07	4	4*1.60	1.60
-18.17	4	4*1.70	1.70
-18.37	6 **(100)**	4*1.90, 2*1.40	1.73

G-Allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	18.4/10 1.84	-62/10 -6.20
23	37.9/23 1.65	-176.2/23 -7.66
62	110.6/62 1.78	-571.72/62 -9.22
100	180.7/100 1.80	-1181.66/100 -11.82

A-Allele

Opening	#LM	Barrier	Avg. Barrier
-2.00	1	1.50	1.50
-2.20	1	1.40	1.40
-2.30	1	1.90	1.90
-2.60	2	2*1.40	1.40
-2.70	1	1.40	1.40
-3.00	1	1.80	1.80
-3.10	1	1.90	1.90
-3.30	5 **(13)**	2.10, 1.80, 3*1.40	1.62
-3.40	1	1.40	1.40
-3.70	1	1.40	1.40
-3.80	2	2.60, 1.90	2.25
-3.90	2	2*1.90	1.90
-4.00	4 **(23)**	2*1.50, 2*1.40	1.45
-4.10	3	2*1.50, 1.40	1.47
-4.30	2	2*1.40	1.40
-4.40	2	2*1.40	1.40
-4.50	1	1.50	1.50
-5.00	1	1.40	1.40
-5.10	3	1.70, 2*1.50	1.57
-5.40	32 **(67)**	16*1.50, 16*1.40	1.45
-5.50	4	4*1.40	1.40
-5.70	2	1.60, 1.50	1.55
-5.80	31 **(104)**	4.60, 3*2.60, 2*2.50, 2.20, 3*2.10, 2.00, 3*1.90, 4*1.80, 3*1.70, 4*1.60, 4*1.50, 2*1.40	1.97

A-Allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
13	20.8/13 1.60	-37/13 -2.85
23	37.7/23 1.52	-75.5/23 -3.28
67	101.7/67 1.52	-302.8/67 -4.52
104	171.5/104 1.65	-516/104 -4.96

4. HLA_G - (L = 386)

NM_002127.5 vs. rs1063320
miRNA: miR-148a-3p
SNP Pos. 233
Target site: 221 - 239

RNAsubopt and Barrier Setting

Offset	Barrier	C-allele	G-allele	C-allele	G-allele
4.0	1.4	9,718,256	8,406,185	11,957	10,473

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
C-allele	-23.90	-4.55	-9.28	-14.62
G-allele	-30.50	-3.24	-11.01	-19.49

Local Minima

MFE	**C-allele**		**G-allele**	
#Pairings	9		11	
Opening Energy	-12.30		-13.10	
Structure Energy	-86.70		-87.50	
MFE Barrier	4.00		4.00	
Minima				
Less Pairings	18.8%	2,248	21.0%	2,201
Equal Pairings	46.9%	5,613	55.9%	5,853
Greater Pairings	34.3%	4,095	23.1%	2,418
Identical to MFE	38.1%	4,554	45.0%	4,714
Minima (- MFE)	**11,956**		**10,472**	
Avg. Opening of All Local Minima (excluding MFE)	**-10.24**		**-11.04**	

Less/Equal Pairings than MFE Target Site

#Pairings	#Minima C-allele		Target Site Approx. Avg.	#Minima G-allele		Target Site Approx. Avg.
6	3.00%	359	-6.88	0.72%	75	-6.30
7	7.89%	943	-5.15	7.26%	760	-5.79
8	7.91%	946	-7.19	7.65%	801	-8.08
9	**46.95%**	**5,613**	**-10.96**	1.48%	155	-7.49
10	**2.18%**	**261**	**-4.80**	3.92%	410	-7.29
11	**0.92%**	**110**	**-6.56**	**55.89%**	**5,853**	**-11.78**
Total Minima	**68.85%**	**8,232**	**-9.43**	**76.91%**	**8,054**	**-10.48**
Avg. Barrier	**1.71**			**1.71**		

Local Minima with Less Pairings

	C-allele	G-allele
#LM Less Pairings:	18.8% 2,248	21.0% 2,201
Avg. Target Site Energy	-6.28	-7.04
Less Pairings Avg. Barrier Height	1.75	1.77

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	C-allele	G-allele
Less	18.8% 6,448	48.8% 5,112
Equal	35.7% 4,272	41.7% 4,372
Less or Equal	**89.7% 10,720**	**90.6% 9,484**

N+ Test

C-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-86.00	33 **(112)**	4*2.30, 4*3.30, 2*3.00, 8*2.80, 2*2.70, 8*2.50, 2.20, 4*1.40	2.55	16*-12.30, 4*-11.60, 8*-6.90, 2*-4.80, 3*-4.10	-9.71
-86.10	24 **(79)**	3.40, 2.90, 2*2.70, 4*2.50, 8*2.30, 8*1.40	2.14	12*-12.30, 8*-6.90, 2*-4.80, 2*-4.10	-9.19
-86.20	17	3*3.50, 3.30, 2*3.00, 2*2.90, 8*2.30, 2.70	2.75	12*-12.30, 3*-4.80, 2*-4.10	-10.01
-86.30	13	2*3.60, 2*3.30, 2.90, 8*2.80	3.00	4*-12.30, 4*-6.90, 2*-4.80, 3*-4.10	-7.59
-86.40	11 **(25)**	4*3.70, 3.30, 2*3.00, 4*2.50	3.10	8*-12.30, -4.80, 2*-4.10	-10.13
-86.50	10 **(14)**	2*3.80, 4*2.30, 4*1.40	2.24	8*-12.30, -4.80, -4.10	-10.73
-86.60	1	3.90	3.90	-4.10	-4.10
-86.70	3	4.00, 2*3.00	3.33	3*-12.30	-12.30
MFE: -86.70	1	4.00	4.00	-12.30	-12.30
-9740.2 Avg. -86.20	**113**	**295.5**	**2.62**	**1081.8**	**-9.68**

C-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
112	-9653.8/112 -86.19	291.5/112 2.60	-1069.5/112 -9.55
79	-6815.8/79 -86.28	207.5/79 2.63	-749.2/79 -9.48
25	-2162.1/25 -86.48	70.4/25 2.82	-259.7/25 -10.39
14	-1211.7/14 -86.55	36.3/14 2.59	-148.3/14 -10.59

N+ Test

G-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-86.80	30 **(118)**	3*3.30, 2.90, 2*2.70, 4*2.30, 8*2.50, 8*2.80, 4*1.40	2.51	16*-13.10, -11.60, -11.20, 8*-7.70, 2*-5.50, 2*-4.80	-10.49
-86.90	27	3*3.40, 3.30, 2*2.90, 2.70, 4*2.50, 8*2.30, 8*1.40	2.28	12*-13.10, -11.60, -11.20, 8*-7.70, 3*-5.50, 2*-4.80	-9.54
-87.00	27 **(61)**	6*3.50, 2*3.30, 2*3.00, 2.90, 8*2.30, 8*1.40	2.45	20*-13.10, -11.60, -11.20, 2*-5.50, 3*-4.80	-11.49
-87.10	12	3*3.60, 3.30, 8*2.80	3.04	4*-13.10, -11.60, 4*-7.70, -5.50, 2*-4.80	-9.16
-87.20	10 **(22)**	4*3.70, 2*3.00, 4*2.50	3.08	8*-13.10, -5.50, -4.80	-11.51
-87.30	9 **(12)**	3.80, 4*2.30, 4*1.40	2.07	8*-13.10, -4.80	-12.18
-87.50	3	4.00, 2*3.00	3.50	3*-13.10	-13.10
MFE: -87.50	1	4.00	4.00	-13.10	-13.10
-10352.2 Avg. -86.99	**119**	**303.00**	**2.55**	**1279.5**	**-10.75**

G-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
118	-10264.7/118 -86.99	299/118 2.53	-1266.4/118 -10.73
61	-5314.4/61 -87.12	162/61 2.66	-684.1/61 -11.21
22	-1920.2/22 -87.28	59.4/22 2.70	-264.22/22 -12.00
12	-1848.2/12 -87.35	29.13/12 2.42	148.9/12 12.41

Ordered by Deepest Local Minima

C-allele

Barrier	#LM	Opening	Avg. Opening
4.00	1	-12.30	-12.30
3.90	1	-4.10	-4.10
3.80	2	-4.10, -4.80	-4.45
3.70	4	-4.10, -4.80, 2*-12.30	-8.38
3.60	2 **(10)**	-4.10, -4.80	-4.45
3.50	3	-4.80, 2*-12.30	-9.80
3.40	1	-4.10	-4.10
3.30	8 **(22)**	3*-4.10, 3*-4.80, 2*-11.60	-6.24
3.20	16	6*-4.10, -4.80, 9*-12.30	-8.76
3.10	11	4*-4.10, 4*-4.0, -7.20, 2*-18.00	-6.87
3.00	37 **(86)**	-4.10, 4*-4.80, -7.20, 4*-9.80, 6*-11.60, 19*-12.30, 2*-18.00	-11.05
2.90	30 **(116)**	6*-4.10, 5*-4.80, 4*-6.30, 3*-7.20, 12*-12.30	-8.10

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	37.5/10 3.75	-67.7/10 -6.77
22	77.8/22 3.54	-151.1/22 -6.87
86	274.1/86 3.19	-779/86 -9.06
116	361.10/116 3.11	-1022/116 -8.81

G-allele

Barrier	#LM	Opening	Avg. Opening
4.00	1	-13.10	-13.10
3.80	1	-4.80	-4.80
3.70	4	-4.80, -5.50, 2*-13.10	-9.13
3.60	3 **(9)**	-4.80, -5.50, -11.60	-7.30
3.50	6	-4.80, -5.50, -11.20, -11.60, 2*-13.10	-9.88
3.40	3	-5.50, -11.20, -11.60	-9.43
3.30	7 **(25)**	3*-4.80, 2*-5.50, -11.20, -11.60	-6.88
3.20	15	-4.80, -5.50, -11.20, 2*-12.30, 9*-13.10, -15.71	-11.98
3.10	8	6*-4.80, -5.50, -15.71	-6.25
3.00	33 **(81)**	4*-4.80, 4*-5.50, -6.80, 2*-12.30, 19*-13.10, 3*-15.71	-11.17
2.90	40 **(121)**	5*-4.80, 7*-5.50, -6.80, 4*-7.00, -7.50, 4*-12.30, 12*-13.10, 4*-15.71, 2*-15.22	-10.11

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
9	33.4/9 3.71	-76.3/9 -8.48
25	87.7/25 3.51	-212.1/25 -8.48
81	259.5/81 3.20	-810.45/81 -10.01
121	375.50/121 3.10	-1214.93/121 -10.04

Ordered by Opening Energy

C-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-3.10	4	2*1.60, 2*1.50	1.55
-3.30	29 **(33)**	2.10, 2.00, 3*1.90, 4*1.80, 4*1.70, 4*1.60, 4*1.50, 8*1.40	1.63
-4.00	21 **(65)**	2.00, 1.90, 3*1.80, 4*1.70, 4*1.60, 4*1.50, 4*1.40	1.62
-4.10	495 **(560)**	81*1.40, 112*1.50, 39*1.60, 61*1.70, 29*1.80, 30*1.90, 17*2.00, 14*2.10, 29*2.20, 11*2.30, 10*2.40, 9*2.50, 8*2.60, 19*2.70, 2.80, 6*2.90, 3.00, 4*3.10, 6*3.20, 3*3.30, 3.40, 3.60, 3.70, 3.80, 3.90	1.84

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	17.8/10 1.78	-32.2/10 -3.22
33	53.6/33 1.62	-108.1/33 -3.28
65	103.1/65 1.59	-237.2/65 -3.65
560	1014/560 1.81	-2274.4/112 -4.06

G-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-3.80	4	2*1.50, 2*1.40	1.45
-4.00	21	2.00, 1.90, 3*1.80, 4*1.70, 4*1.60, 4*1.50, 4*1.40	1.62
-4.70	17	1.90, 1.80, 3*1.70, 4*1.60, 4*1.50, 4*1.40	1.58
-4.80	418	3.80, 3.70, 3.60, 3.50, 3*3.30, 3.20, 6*3.10, 4*3.00, 5*2.90, 2*2.80, 10*2.70, 10*2.60, 10*2.50, 7*2.40, 10*2.30, 26*2.20, 14*2.10, 14*2.00, 24*1.90, 23*1.80, 58*1.70, 32*1.60, 88*1.50, 67*1.40	1.85

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	16.8/10 1.68	-39.2/10 -3.92

25	39.9/25 1.60	-99.2/25 -3.97
60	125/60 2.08	-265.5/60 -4.43
460	840.10/460 1.83	-2185.50/460 -4.75

5. PARP1 (L = 769)

NM_001618.3 vs. rs8679
miRNA: miR-145-5p
SNP Pos. 607
Target site: 592-614

RNAsubopt and Barrier Setting

Offset	Barrier	U-allele	C-allele	U-allele	C-allele
6.0	1.4	10,610,542	1,582,333	14,281	1,709

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
U-allele	-19.70	0.93	-12.19	-7.51
C-allele	-19.20	-0.65	-20.90	1.70

Local Minima

MFE	**U-allele**	**C-allele**
#Pairings	11	22
Opening Energy	-18.10	-31.12
Structure Energy	-186.90	-188.70
MFE Barrier	3.00	3.00
Minima		
Less Pairings	79.1% 11,295	12.3% 210
Equal Pairings	7.9% 1,226	87.7% 1,499
Greater Pairings	12.3% 1,760	0.0% 0
Identical to MFE	8.6% 1,226	42.2% 807
Minima (- MFE)	**14,281**	**1,709**
Avg. Opening of All Local Minima (excluding MFE)	**-12.83**	**-30.81**

Less/Equal Pairings than MFE Target Site

#Pairings	#Minima U-allele	Target Site Approx. Avg.	#Minima C-allele	Target Site Approx. Avg.
6	52.06% 7,434	-8.75	-	-
7	21.70% 3,099	-10.19	-	-
10	5.34% 762	-17.90	-	-
11	**8.58% 1,226**	**-18.10**	-	-
12	**0.01% 1**	**-19.20**	-	-
17	**0.01% 2**	**-22.40**	-	-
20	**1.56% 223**	**-27.98**	12.17% 208	-30.46
21	**0.01% 2**	**-25.50**	0.12% 2	-27.40
22	**10.73% 1,532**	**-28.97**	**87.71% 1,499**	**-30.87**
Total Minima	100% 14,281	-12.83	100% 1,709	-30.81

Local Minima with Less Pairings

	U-allele	C-allele
#LM Less Pairings:	79.1% 11,295	12.29% 210
Avg. Target Site Energy	-9.76	-30.43
Less Pairings Avg. Barrier Height	1.42	1.51

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	U-allele	C-allele
Less	79.1% 11,295	12.3% 902
Equal	8.6% 1,226	42.2% 807
Less or Equal	**87.7% 12,521**	**100% 1,709**

N+ Test

U-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-186.30	48 **(103)**	24*2.40, 10*2.00, 10*1.70, 4*1.50	2.10	12*-28.72, 10*-18.10, 2*-17.90, 8*-10.20, 16*-8.80	-16.33
-186.40	14 **(55)**	12*2.50, 2*2.00	2.43	8*-18.10, 6*-17.90	-18.01
-186.50	18	10*2.60, 4*2.40, 4*2.00	2.42	4*-29.22, 2*-18.10, 4*-10.20, 8*-8.80	-14.68
-186.60	12 **(23)**	6*2.70, 2*2.60, 2*2.00, 2*1.20	2.32	4*-29.22, 6*-18.10, 2*-17.90	-12.77
-186.70	6 **(11)**	2*2.80, 4*2.40	2.53	4*29.22, 2*-17.90	-25.45
-186.80	4	2.90, 2.70, 2*2.60	2.70	4*-29.22	-29.22
-186.90	1	3.00	3.00	-18.10	-18.10
MFE: -186.90	1	3.00	3.00	-18.10	-18.10
-19389.4 Avg. -186.44	**104**	**238**	**2.29**	**1867.36**	**-17.96**

U-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
103	-19202.5/103 -186.43	235/103 2.28	-1849.26/103 -17.95
55	-1065.42/55 -186.55	134.4/55 2.44	-1065.42/55 -19.37
23	-4293.5/23 -186.67	56.8/23 2.47	-548.94/23 -23.87
11	-2054.3/11 -186.75	29/11 2.64	-287.66/11 -26.15

N+ Test

C-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-187.90	40 **(123)**	20*2.20, 12*2.10, 4*1.70, 4*1.50	2.05	4*-30.52, 12*-30.62, 24*-31.12	-30.91
-188.00	32 **(83)**	20*2.30, 4*2.10, 4*1.70, 4*1.50	2.10	4*-30.52, 16*-31.12, 12*-30.62	-30.86
-188.10	24	12*2.40, 4*2.10, 4*1.70, 4*1.50	2.08	12*-30.62, 4*-30.52, 8*-31.12	-30.77
-188.20	12 **(27)**	4*2.50, 4*1.70, 4*1.50	1.90	12*-30.62	-30.62
-188.40	4	4*2.40	2.40	4*-31.12	-31.12
-188.50	4 **(11)**	2.80, 2.70, 2*2.60	2.68	4*-31.12	-31.12
-188.60	4	4*2.40	2.40	4*-31.12	-31.12
-188.70	3	2.70, 2*2.60	2.63	3*-31.12	-31.12
MFE: -188.70	1	3.00	3.00	-31.12	-31.12
-23321.6 Avg. -188.08	**124**	**262.8**	**2.12**	**3827.68**	**-30.87**

C-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
123	-23132.9/123 -188.07	259.8/123 2.11	-3796.56/123 -30.87
83	-15616.9/83 -188.16	117.8/83 2.14	-2560.16/83 -30.85
27	-5086.5/27 -188.39	60.6/27 2.24	-834.24/27 -30.90
11	-2074.5/11 -188.6	28.2/11 2.56	342.32/11 -31.12

Ordered by Deepest Local Minima

U-allele

Barrier	#LM	Opening	Avg. Opening
3.00	1	-18.10	-18.10
2.90	1	-29.22	-29.22
2.80	2	2*-17.90	-17.90
2.70	7 **(11)**	4*-18.10, 3*-29.22	-22.87
2.60	14 **(25)**	4*-8.80, 4*-10.20, 2*-18.10, 4*-29.22	-16.36
2.50	12	4*-17.90, 8*-18.10	-18.03
2.40	32 **(69)**	6*-8.80, 6*-10.20, 2*-17.90, 6*-18.10, 4*-28.72, 8*-29.22	-18.97
2.30	54 **(123)**	4*-8.50, 12*8.80, 12*-10.20, 8*-17.90, 6*-18.10, 4*-28.72, 8*-29.22	-15.97

U-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
11	30.4/11 2.76	-243.18/11 -22.11
25	66.8/25 2.67	-472.26/25 -18.89
69	173.6/69 2.52	-1295.7/69 -18.78
123	297.8/123 2.42	-2158.14/123 -17.55

C-allele

Barrier	#LM	Opening	Avg. Opening
2.80	1	-31.12	-31.12
2.70	2	2*-31.12	-31.12
2.60	4	4*-31.12	-31.12
2.50	4 **(11)**	4*-30.62	-30.62
2.40	20 **(31)**	16*-31.12, 4*30.62	-31.02
2.30	20	16*-31.12, 4*-30.62	-31.02
2.20	20 **(71)**	16*-31.12, 4*-30.62	-31.02
2.10	64 **(135)**	16*-30.52, 48*-31.12	-30.97

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
11	28.6/11 2.60	-340.32/11 -30.94
31	76.6/31 2.47	-960.72/31 -30.99
71	166.6/71 2.35	-2201.52/71 -31.01
135	301/135	-4183.60/116

	2.23	-30.99

Ordered by Opening Energy

U-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-8.50	1,300	4*2.30, 6*2.10, 16*2.00, 13*1.90, 48*1.80, 109*1.70, 64*1.60, 161*1.50, 183*1.40, 213*1.30, 483*1.20	1.39
-8.80	6,134	4*2.60, 6*2.40, 12*2.30, 10*2.20, 32*2.10, 172*2.00, 166*1.90, 230*1.80, 568*1.70, 378*1.60, 748*1.50, 760*1.40, 934*1.30, 2114*1.20	1.42

U-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	21.8/10 2.18	-85/10 -8.50
20	41.8/20 2.09	-170/20 -8.50
60	116.3/60 1.94	-510/60 -8.50
100	187/100 1.87	-850/100 -8.50

C-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-27.40	2	2*1.30	1.30
-27.50	4	2*1.40, 2*1.30	1.35
-30.02	24	16*1.50, 8*1.30	1.43
-30.12	40	16*1.50, 8*1.40, 16*1.30	1.40
-30.52	184	40*2.10, 24*1.60, 40*1.50, 44*1.40, 36*1.30	1.58

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	14/10 1.40	-284.88/10 -28.49
20	29/20 1.45	-585.08/20 -29.54
70	98.4/70 1.41	-2090.08/70 -29.86
254	378.8/254 1.49	-7705.76/254 -30.34

6. WFS1 (L = 797)

NM_001145853.1 vs. rs1046322
miRNA: miR-668-3p
SNP Pos. 253
Target site: 234 - 258

RNAsubopt and Barrier Setting

Offset	Barrier	G-allele	A-allele	G-allele	A-allele
2.7	0.8	11,404,145	11,362,245	79,577	79,273

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
G-allele	-26.60	-4.421	-10.641	-15.959
A-allele	-20.60	-1.723	-14.196	-6.404

Local Minima

MFE	**G -allele**	**A-allele**
#Pairings	12	12
Opening Energy	-15.72	-15.72
Structure Energy	-290.70	-290.70
MFE Barrier	2.70	2.70
Minima		
Less Pairings	0.34% 272	0%
Equal Pairings	92.9% 73,943	93.3% 73,943
Greater Pairings	6.74% 5,360	6.7% 5,328
Target Site Identical to MFE	92.9% 73,943	93.3% 73,943
Minima (- MFE)	**79,575**	**79,271**
Avg. Opening of All Local Minima (excluding MFE)	**-15.64**	**-15.67**

Less/Equal Pairings than MFE

#Pairings	#Minima G-allele	Target Site Approx. Avg.	#Minima A-allele	Target Site Approx. Avg.
10	0.34% 272	-7.56	0	-
Total:	**0.34% 272**	**-7.56**	**0**	**-**
Avg. Barrier:	**1.01**		**-**	

Local Minima with Less or Equal Approx. Target Site Energy than the MFE

	G-allele	A-allele
Less	7% 5,600	6.7% 5,328
Equal	92.9% 73,943	93.3% 73,943
Less or Equal	99.96% 79,543	100% 79,271

N+ Test

G-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-290.50	96	16*2.50, 16*2.30, 32*1.90, 32*1.30	1.87	96*-15.72	-15.72
-290.60	64	32*2.60, 16*1.90, 16*1.30	2.10	64*-15.72	-15.72
MFE: -290.70	16	16*2.70	2.70	16*-15.72	-15.72

G-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
10	-290.70	2.70	-15.72
20	-290.68	2.50	-15.72
60	-290.63	2.44	-15.72
100	-290.59	2.15	-15.72

N+ Test

A-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-290.50	96	16*2.50, 16*2.30, 32*1.90, 32*1.30	1.87	96*-15.72	-15.72
-290.60	64	32*2.60, 16*1.90, 16*1.30	2.10	64*-15.72	-15.72
MFE: -290.70	16	16*2.70	2.70	16*-15.72	-15.72

G-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
10	-290.70	2.70	-15.72
20	-290.68	2.50	-15.72
60	-290.63	2.44	-15.72
100	-290.59	2.15	-15.72

Ordered by Deepest Local Minima

G-Allele

Barrier	#LM	Opening	Avg. Opening
2.70	15	-15.72	-15.72
2.60	32	-15.72	-15.72
2.50	16	-15.72	-15.72
2.40	16	-15.72	-15.72
2.30	80	-15.72	-15.72

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	2.70	-15.72
20	2.68	-15.72
60	2.65	-15.72
100	2.50	-15.72

A-Allele

Barrier	#LM	Opening	Avg. Opening
2.70	15	-15.72	-15.72
2.60	32	-15.72	-15.72
2.50	16	-15.72	-15.72
2.40	16	-15.72	-15.72
2.30	80	-15.72	-15.72

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	2.70	-15.72
20	2.68	-15.72
60	2.65	-15.72
100	2.50	-15.72

Ordered by Opening Energy

G-allele

Opening	#LM	Barrier	Avg. Barrier	Structure Energy	Avg. Structure
-7.30	32	32*0.90	0.90	-288.90	-288.90
-7.60	240	32*1.20, 64*1.10, 64*1.00, 80*0.90	1.02	32*-289.20 64*-289.10 64*-289.00 80*-288.90	-289.02

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	0.90	-7.30
20	0.90	-7.30
60	0.90	32*-7.30, 28*-7.60 Avg. -7.44
100	0.90	32*-7.30, 68*-7.60 Avg. -7.50

A-allele

Opening	#LM	Barrier	Avg. Barrier	Structure Energy	Avg. Structure
-14.92	5,328	96*0.80 1,544*0.90 1,400*1.00 1040*1.10 680*1.20 344*1.30 80*1.50 64*1.60 64*1.70 16*1.90	1.06	1544*-288.90 1,400*-289.00 1088*-289.10 552*-289.20 376*-289.40 96*-289.50 96*-289.60 96*-289.70 16*-289.80 64*-289.90	-289.09

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	0.90	-14.92
20	0.90	-14.92
60	0.90	-14.92
100	0.90	-14.92

7. IL23R (L = 851)

NM_144701.2 vs. rs10889677
miRNA: let-7e
SNP Pos. 309

	C-allele	A-allele
Target Site:	291 - 310	291 - 308

RNAsubopt and Barrier Setting

Offset	Barrier	C-allele	A-allele	C-allele	A-allele
2.0	1.2	396,014	756,122	964	1,080

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
C-allele	-24.70	0.00	-25.82	1.12
A-allele	-20.70	-0.05	-21.40	0.70

Local Minima

MFE	C-allele		A-allele	
#Pairings	17		15	
Opening Energy	-25.90		-21.70	
Structure Energy	-228.40		-224.60	
MFE Barrier	2.00		2.00	
Minima				
Less Pairings	0.0%	0	35.6%	384
Equal Pairings	100%	963	64.4%	695
Greater Pairings	0.0%	0	0.0%	0
Identical to MFE	100%	963	64.4%	695
Minima (- MFE)	**963**		**1,079**	
Avg. Opening of All Local Minima (excluding MFE)	**-25.90**		**-18.46**	

#Pairings	#Minima C-allele		Target Site Approx. Avg.	#Minima A-allele		Target Site Approx. Avg.
12	-		-	29.7%	320	-12.50
13	-		-	5.9%	64	-13.10
14	-		-	-		-
15	-		-	**64.4%**	**695**	**-21.70**
16	-		-	-		-
17	**100%**	**963**	**-25.90**	-		-
Total Minima	**100%**	**963**	**-25.90**	**100%**	**1,079**	**-18.46**
Avg. Barrier	**1.37**			**1.44**		

Local Minima with Less Pairings

	C-allele	A-allele
#LM Less Pairings:	0.0%	35.6% 384
Avg. Target Site Energy	-	-12.60
Less Pairings Avg. Barrier Height	-	1.40

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	C-allele		A-allele	
Less	0.0%	0	35.6%	384
Equal	100%	963	64.4%	695
Less or Equal	100%	963	100%	1,079

N+ Test

C-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-228.10	40 **(135)**	40*1.70	1.70	40*-25.90	-25.90
-228.20	32	32*1.70	1.70	32*-25.90	-25.90
-228.30	32 **(63)**	32*1.90	1.90	32*-25.90	-25.90
-228.40	31 **(31)**	31*2.00	2.00	31*25.90	-25.90
MFE -228.40	1	2.00	2.00	-25.90	-25.90
31040.8 Avg. -228.24	**136**	**247.2**	**1.80**	**3522.4**	**-25.90**

C-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
135	-30812.4/135 -228.24	245.2/135 1.86	-3496.5/135 -25.90
63	-14386/63 -228.35	122.8/63 1.95	-1631.7/63 -25.90
31	-7080.4/31 -228.4	62/31 2.00	-802.9/31 -25.90
10	-2284/10 -228.4	20/10 2.00	-259.00 -25.90

N+ Test

A-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-224.30	40 **(135)**	40*1.70	1.70	40*-21.70	-21.70
-224.40	32	32*1.70	1.70	32*-21.70	-21.70
-224.50	32 **(63)**	32*1.90	1.90	32*-21.70	-21.70
-224.60	31 **(31)**	31*2.00	2.00	31*-21.70	-21.70
MFE: -224.60	1	2.00	2.00	-21.70	-21.70
-30524 Avg. -224.44	**136**	**247.2**	**1.82**	**2951.2**	**-21.70**

A-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
135	-30299.4/135 -224.44	245.2/135 1.82	-2929.5/135 -21.70
63	-14146.6/63 -224.55	122.8/63 1.95	-1367.1/63 -21.70
31	-6962.6/31 -224.60	62/31 2.00	672.7/31 -21.70
10	-2246/10 -224.60	20/10 2.00	-217/10 -21.70

Ordered by Deepest Local Minima

C-allele

Barrier	#LM	Opening	Avg. Opening
2.00	31	31*-25.90	-25.90
1.90	32	32*-25.90	-25.90
1.80	72	72*-25.90	-25.90

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	20/10 2.00	-259/10 -25.90
20	40/20 2.00	-518/20 -25.90
60	117.1/60 1.95	-1554/69 -25.90
100	189.4/100 1.89	-2590/100 -25.90

A-allele

Barrier	#LM	Opening	Avg. Opening
2.00	31	31*-21.70	-21.70
1.90	32	32*-21.70	-21.70
1.70	72	72*-21.70	-21.70

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	20/10 2.00	-217/10 -21.70
20	40/20 2.00	-434/20 -30.99
60	117.1/60 1.95	-1302/60 -21.70
100	185.7/100 1.86	-2170/100 -21.70

Ordered by Opening Energy

C-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-25.90	963	31*2.00, 32*1.90, 72*1.70, 56*1.50, 232*1.40, 200*1.30, 340*1.20	1.37

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	20/10 2.00	-259/10 -25.90
20	40/20 2.00	-518/20 -25.90
60	117.1/60 1.95	-1554/60 -25.90
100	185.7/100 1.86	-2590/100 -25.90

A-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-12.50	320	128*1.30, 64*1.40, 64*1.50, 64*1.60	1.42

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	16/10 1.60	-125/10 -12.50
20	32/20 1.60	-250/20 -12.50

60	72/60 1.60	-750/60 -12.50
100	156.4/100 1.56	-1250/100 -12.50

8. RYR3 (L = 880)

NM_001036.3 vs. rs1044129
miRNA: miR-367
SNP Pos: 839

	A-allele	G-allele
Target Site:	835 - 857	830 - 857

RNAsubopt and Barrier Setting

Offset	Barrier	A-allele	G-allele	A-allele	G-allele
2.3	1.2	13,569,252	5,006,010	19,936	2,628

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
A-allele	-14.80	-1.22	-11.42	-3.38
G-allele	-15.30	-0.19	-19.65	4.35

Local Minima

MFE	A-allele		G-allele	
#Pairings	7		14	
Opening Energy	-10.85		-11.60	
Structure Energy	-186.60		-188.40	
MFE Barrier	2.30		2.30	
Minima				
Less Pairings	1.6%	324	0.0%	0
Equal Pairings	56.3%	11,227	96.6%	2,537
Greater Pairings	42.1%	8,384	3.4%	90
Identical to MFE	56.2%	11,203	90.6%	2,379
Minima (- MFE)	**19,935**		**2,627**	
Avg. Opening of All Local Minima (excluding MFE)	**-10.93**		**-11.75**	

Local Minima

#Pairings	#Minima A-allele		Target Site Approx. Avg.	#Minima G-allele	Target Site Approx. Avg.
6	1.6%	324	-9.19	-	-
7	**56.3%**	**11,227**	**-10.85**	-	-
8	**36.7%**	**7,322**	**-11.09**	-	-
9	**5.3%**	**1,062**	**-11.28**	-	-
10	-		-	-	-
11	-		-	-	-
12	-		-	-	-
13	-		-	-	-
14	-		-	**2,537**	**-11.57**
Total Minima	**19,935**		**-10.93**	**2,537**	**-11.57**
Avg. Barrier	**1.42**			**1.44**	

Local Minima with Less Pairings

	A-allele	G-allele
#LM Less Pairings:	1.6% 324	0
Avg. Target Site Energy	-9.19	-
Less Pairings Avg. Barrier Height	1.37	-

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	A-allele		G-allele	
Less	19.1%	3,816	5.3%	140
Equal	56.2%	11,203	90.6%	2,379
Less or Equal	**73.3%**	**15,019**	**95.9%**	**2,519**

N+ Test

A-allele - 100 Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-186.40	90 **(185)**	24*2.10, 12*2.00, 12*1.80, 12*1.50, 30*1.40	1.73	30*-11.58, 60*-10.85	-11.09
-186.50	72 **(95)**	24*2.20, 8*2.00, 8*1.80, 8*1.50, 24*1.40	1.79	12*-11.58, 60*-10.85	-10.97
-186.60	23 **(23)**	15*2.30, 8*1.40	1.99	23*-10.85	-10.85
MFE: -186.60	1	2.30	2.30	-10.85	-10.85
-34682.4 Avg. -186.46	**186**	**332.8**	**1.79**	**2048.76**	**-11.01**

A-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
10	-1866/10 -186.60	23/10 2.30	-108.5/10 -10.85
23	-4291.8/23 -186.60	45.7/23 1.99	-249.55/23 -10.85
95	-17719.8/95 -186.52	174.5/95 1.84	-1039.5/95 -10.94
185	-34495.8/185 -186.46	330.5/185 1.79	-2037.9/185 -11.02

N+ Test

G-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-188.00	62 **(139)**	34*1.90, 4*1.80, 4*1.60, 20*1.40	1.71	62*-11.60	-11.60
-188.10	48 **(77)**	24*2.00,4*1.80, 4*1.60, 16* 1.40	1.75	48*-11.60	-11.60
-188.20	18 **(29)**	8*2.10, 4*1.60, 6*1.40	1.76	18*-11.60	-11.60
-188.30	6	4*2.20, 2*1.40	1.93	6*-11.60	-11.60
-188.40	5	3*2.30, 2*1.40	1.94	5*-11.60	-11.60
MFE: -188.40	1	2.30	2.30	-11.60	-11.60
-26332.6 Avg. -188.09	**140**	**245.4**	**1.75**	**-1624**	**-11.60**

G-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
11	-2071.8/11 -188.35	21.3/11 1.94	127.6/11 -11.60
29	-5459.4/29 -188.26	52.9/29 1.82	-336.4/29 -11.60
77	-14488.2/77 -188.16	136.9/77 1.78	-893.2/77 -11.60
139	-26144.2/139 -188.09	243.1/139 1.75	-1612.4/139 -11.60

Ordered by Deepest Local Minima

A-allele

Barrier	#LM	Opening	Avg. Opening
2.30	15	15*-10.85	-10.85
2.20	24	24*-10.85	-10.85
2.10	24	24*-10.85	-10.85
2.00	136	104*-10.85, 32*-11.58	-11.02

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	23/10 2.30	-108.5/10 -10.85
20	45.5/20 2.28	-217/20 -10.85
63	137.7/63 2.19	-683.55/63 -10.85

199	409.7/100 2.06	-2182.51/100 -10.97

G-allele

Barrier	#LM	Opening	Avg. Opening
2.30	3	3*-11.60	-11.60
2.20	4	4*-11.60	-11.60
2.10	8	8*-11.60	-11.60
2.00	24	24*-11.60	-11.60
1.90	34	34*-11.60	-11.60
1.80	62	62*-11.60	-11.60

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	22/10 2.20	-116/10 -11.60
20	42.5/20 2.13	-232/20 -11.60
73	145.1/60 1.99	-846.8/73 -11.60
135	256.7/100 1.90	-1566/135 -11.60

Ordered by Opening Energy

A-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-8.39	24	24*1.30	1.30
-9.19	324	16*1.60, 40*1.50, 88*1.40, 180*1.30	1.37

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	13/10 1.30	-83.9/10 -8.39
20	26/20 1.30	-167.8/20 -8.39
60	86.8/60 1.45	-532.2/60 -8.87
100	144.8/100 1.45	-899.8/100 -9.00

G-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-11.00	140	4*1.70, 4*1.60, 12*1.50, 58*1.40, 62*1.30	1.38

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	16.2/10 1.62	-110/10 -11.00
20	31.2/20 1.56	-220/20 -11.00
60	87.2/60 1.45	-660/60 -11.00
100	141/100 1.41	-1100/100 -11.00

9. AGTR1 (L = 888)

NM_032049.3 vs. rs5186
miRNA: miR-155-5p
SNP Pos: 86

	A-allele	C-allele
Target Site:	57 - 90	57 - 90

RNAsubopt and Barrier Setting

Offset	Barrier	A-allele	C-allele	A-allele	C-allele
2.3	0.8	546,284	2,147,815	2,309	14,944

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
A-allele	-20.90	-0.577	-28.561	7.661
C-allele	-16.60	-2.268	-27.879	11.279

Local Minima

MFE	A -allele	C-allele
#Pairings	22	24
Opening Energy	-18.60	-21.30
Structure Energy	-188.00	-188.80
MFE Barrier	2.30	2.30
Minima		
Less Pairings	8.1% 186	0%
Equal Pairings	91.9% 2,122	100% 14,943
Greater Pairings	0%	0%
Target Site Identical to MFE	45.8% 1,058	25% 3,735
Minima (- MFE)	**2,308**	**14,943**
Avg. Opening of All Local Minima (excluding MFE)	**-18.54**	**-21.30**

Less/Equal Pairings than MFE

#Pairings	#Minima A-allele	Target Site Approx. Avg.	#Minima C-allele	Target Site Approx. Avg.
18	8.1% 186	-17.90	0	-
22	91.9% 2,122	-18.60	0	-
24	0	-	100% 14,943	-21.30
Total:	**2,308**	**-18.54**	**14,943**	**-21.30**
Avg. Barrier:	**1.01**		**1.00**	

Local Minima with Less Pairings

	A-allele	C-allele
#LM Less Pairings:	8.1% 186	0%
Avg. Target Site Energy	-17.90	-
Less Pairings Avg. Barrier Height	0.97	-

Local Minima with Less or Equal Approx. Target Site Energy than the MFE

	A-allele	C-allele
Less	8.3% 191	0%
Equal	91.7% 2,117	100% 14,943
Less or Equal	100% 2,308	100% 14,943

N+ Test

A-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-187.30	44 **(122)**	30*1.60, 10*1.40, 4*0.90	1.49	44*-18.60	-18.60
-187.31	2	2*1.61	1.61	2*-18.60	-18.60
-187.40	22 **(76)**	10*1.70, 10*1.40, 2*0.90	1.49	22*-18.60	-18.60
-187.41	6	3*1.71, 2*1.60	1.39	6*-18.60	-18.60
-187.50	14	10*1.80, 2*1.60, 2*1.40	1.71	14*-18.60	-18.60
-187.51	2	2*1.81	1.81	2*-18.60	-18.60
-187.60	10	4*1.90, 6*1.40	1.60	110*-18.60	-18.60
-187.61	2	2*1.91	1.91	2*-18.60	-18.60
-187.70	8 **(20)**	4*2.00, 2*1.60, 2*1.40	1.75	8*-18.60	-18.60
-187.71	2	2*2.01	2.01	2*-18.60	-18.60
-187.80	4 **(10)**	2*2.10, 2*1.40	1.75	4*-18.60	-18.60
-187.90	4	2*2.20, 2*1.40	1.80	4*-18.60	-18.60
-188.00 MFE	2	2*2.30	2.30	2*-18.60	-18.60
Total: -22870.14 Avg. -187.46	**122**	**194.21**	**1.59**	**-2269.2**	**-18.60**

A-allele - N+ Averages

Test #LM-MFE	Avg. Structure Energy	Avg. Barrier	Avg. Opening Energy
10	-1878.8/10 -187.88	18.8/10 1.88	-18.60
20	-3755.82/20 -187.79	36.82/20 1.84	-18.60
76	-14254.32/76 187.56	125.39/76 1.65	-18.60
122	-22870.14/122 -187.46	194.21/122 1.59	-18.60

N+ Test

C-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-188.50	48 **(120)**	16*2.00, 8*1.80, 12*1.60, 12*1.40	1.72	48*-21.30	-21.30
-188.51	12	8*2.01, 4*1.80	1.94	12*-21.30	-21.30
-188.60	24 **(60)**	8*2.10, 4*1.80, 12*1.40	1.70	24*-21.30	-21.30
-188.70	24 **(36)**	8*2.20, 4*1.80, 12*1.40	1.73	36*-21.30	-21.30
-188.80 MFE	12 **(12)**	8*2.30, 4*1.80	2.13	12*-21.30	-21.30
Total: -22630.92 Avg. -188.59	**120**	**191.4/120**	**1.60**	**-2556/120**	**-21.30**

C-allele - N+ Averages

Test #LM-MFE	Avg. Structure Energy	Avg. Barrier	Avg. Opening Energy
12	-2265.6/12 -188.80	26.6/12 2.13	-21.30
36	-6794.4/36 -188.73	68.2/36 1.89	-21.30
60	-11320.8/60 -188.68	109/60 1.82	-21.30
120	-22630.92/120 -188.59	191.4/120 1.60	-21.30

Ordered by Deepest Local Minima

A-Allele

Barrier	#LM	Opening	Avg. Opening
2.30	2	2*-18.60	-18.60
2.20	2	2*-18.60	-18.60
2.10	2	2*-18.60	-18.60
2.01	2	2*-18.60	-18.60

2.00	4 **(12)**	4*-18.60	-18.60
1.91	2	2*-18.60	-18.60
1.90	4	4*-18.60	-18.60
1.81	2 **(20)**	2*-18.60	-18.60
1.80	10	10*-18.60	-18.60
1.71	4	4*-18.60	-18.60
1.70	10	10*-18.60	-18.60
1.61	2	2*-18.60	-18.60
1.60	36 **(82)**	34*-18.60, 2*-17.90	-18.56
1.51	10	10*-18.60	-18.60
1.50	30 **(122)**	28*-18.60, 2*-17.90	-18.55

A-Allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
12	25.22/12 2.10	-18.60
24	40.26/20 2.01	-18.60
83	125.92/82 1.54	-18.58
170	186.02/122 1.52	-18.58

C-Allele

Barrier	#LM	Opening	Avg. Opening
2.30	8	-21.30	-21.30
2.20	8 **(16)**	-21.30	-21.30
2.10	8 **(24)**	-21.30	-21.30
2.01	8	-21.30	-21.30
2.00	16	-21.30	-21.30
1.91	8	-21.30	-21.30
1.90	20 **(76)**	-21.30	-21.30
1.81	8	-21.30	-21.30
1.80	92 **(176)**	-21.30	-21.30

C-Allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
16	36/16 2.25	-21.30

24	52.8/24 2.20	-21.30
76	154.16/76 2.03	-21.30
176	334.24/176 1.90	-21.30

Ordered by Opening Energy

A-Allele

Opening	#LM	Barrier	Avg. Barrier
-17.30	5	1.00, 2*0.90, 2*0.80	0.88
-17.90	186	2*1.60, 2*1.50, 6*1.40, 2*1.31, 8*1.30, 2*1.21, 10*1.20, 2*1.11, 14*1.10, 6*1.01, 20*1.00, 2*0.91, 44*0.90, 10*0.81, 26*0.80	2.13

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	1.19	-17.60
20	1.28	-17.90
60	1.20	-17.90
100	1.10	-17.90

C-Allele

Opening	#LM	Barrier	Avg. Barrier
-21.30	14,943	8*2.30, 8*2.20, 8*2.10, 8*2.01, 16*2.00, 8*1.91, 20*2.00, 8*1.81, 92*1.80, 24*1.71, 76*1.70, 16*1.61, 196*1.60, 52*1.51, 132*1.50, 64*1.41, 704*1.40, 104*1.31, 520*1.30, 96*1.21, 600*1.20, 204*1.11, 1696*1.10, 220*1.01, 2204*1.00, 428*0.91, 3092*0.90, 592*0.81, 3748*0.80	1.00

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	2.15	-21.30
20	1.94	-21.30
60	1.80	-21.30
100	1.79	-21.30

10. FGF20 (L = 903)

NM_019851.2 vs. rs12720208
miRNA: miR-433-3p
SNP Pos. 182

	C-allele	U-allele
Target Site:	166 - 187	166 - 187

RNAsubopt and Barrier Setting

Offset	Barrier	C-allele	U-allele	C-allele	U-allele
2.3	0.8	1,213,806	2,514,455	5,937	7,784

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
C-allele	-14.50	-2.356	-9.274	-5.226
U-allele	-12.30	-4.655	-10.123	-2.177

Local Minima

MFE	**C -allele**	**U-allele**
#Pairings	9	10
Opening Energy	-4.10	-7.30
Structure Energy	-159.50	-158.90
MFE Barrier	2.30	2.30
Minima		
Less Pairings	0.6% 34	7.5% 582
Equal Pairings	51.3% 3,048	92.5% 7,201
Greater Pairings	48.1% 2,854	0%
Target Site Identical to MFE	50.9% 3,020	89% 6,927
Minima (- MFE)	**5,936**	**7,783**
Avg. Opening of All Local Minima (excluding MFE)	**-6.40**	**-8.92**

Less/Equal Pairings than MFE

#Pairings	#Minima C-allele	Target Site Approx. Avg.	#Minima U-allele	Target Site Approx. Avg.
6	0.1% 6	-5.80	1.7% 132	-5.80
7	0.5% 28	-6.80	5.8% 450	-6.79
9	51.3% 3,048	-4.10	0	-
10	**30.4% 1,804**	**-7.66**	**92.5% 7,201**	**-9.11**
Total:	**82.3% 4,886**	**-5.43**	**100% 7,783**	**-8.92**
Avg. Barrier:	**0.97**		**1.05**	

Local Minima with Less Pairings

	C-allele	U-allele
#LM Less Pairings:	0.6% 34	7.5% 582
Avg. Target Site Energy	-6.62	-6.57

Less Pairings Avg. Barrier Height	0.88	1.03

Local Minima with Less or Equal Approx. Target Site Energy than the MFE

	C-allele		U-allele	
Less	0.7%	42	11%	856
Equal	50.9%	3,020	44.5%	3,463
Less or Equal	51.6%	3,062	55.5%	4,319

N+ Test

C-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-158.80	70 **(163)**	54*1.60, 1.40, 9*1.20, 6*0.90	1.49	9*-11.10, 3*-10.60, 2*-8.10, 2*-7.30, -4.30, 53*-4.10	-5.49
-158.90	45 **(93)**	36*1.70, 1.60, 5*1.20, 3*0.90	1.59	4*-11.10, 2*-10.60, -8.10, -7.30, 37*-4.10	-5.17
-159.00	23	18*1.80, 3*1.20, 2*0.90	1.64	2*-11.10, -10.60, 20*-4.10	-4.99
-159.10	12 **(25)**	9*1.90, 2*1.20, 0.90	1.70	-11.10, 11*-4.10	-4.68
-159.20	7 **(13)**	6*2.00, 1.20	1.88	7*-4.10	-4.10
-159.30	3	3*2.10	2.10	3*-4.10	-4.10
-159.40	2	2*2.20	2.20	2*-4.10	-4.10
MFE: -159.50	1	2.30	2.30	-4.10	-4.10
-25903.3 Avg. -158.92	**163**	**259.9**	**1.59**	**-841.1**	**-5.16**

C-allele - N+ Averages

Test #LM-MFE	Avg. Structure Energy	Avg. Barrier	Avg. Opening Energy
13	-2070.6/13 -159.28	26.2/13 2.02	-53.3/13 -4.10
25	-3979.8/25 -159.19	46.6/25 1.86	-109.5/25 -4.38
93	-14787.3/93 -159.00	155.9/93 1.68	-457/93 -4.91
163	-25903.3/163 -158.92	259.9/163 1.59	-841.1/163 -5.16

N+ Test

U-allele Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-158.50	51 **(101)**	37*1.80, 4*1.30, 6*1.20, 4*0.90	1.62	25*-11.10, 25*-7.30, -6.80	-9.15
-158.60	28 **(50)**	18*1.90, 2*1.80, 2*1.30, 4*1.20, 2*0.90	1.68	14*-11.10, 14*-7.30	-9.20
-158.70	16 **(22)**	14*2.00, 2*1.20	1.90	8*-11.10, 8*-7.30	-9.20
-158.80	4	4*2.20	2.20	2*-11.10, 2*-7.30	-9.20
MFE: -158.90	2	2*2.30	2.30	-11.10, -7.30	-9.20
-16016.5 Avg. -158.58	**101**	**173.4**	**1.71**	**926.8**	**-9.20**

U-allele - N+ Averages

Test #LM-MFE	Avg. Structure Energy	Avg. Barrier	Avg. Opening Energy
20	-3492.2/22 -158.74	43.8/22 1.99	202.4/22 -9.20
48	-7933/50 -158.66	90.8/50 1.82	460/50 -9.20
101	-16016.5/101 -158.58	173.4/101 1.71	926.8/101 -9.20

Ordered by Deepest Local Minima

C-allele

Barrier	#LM	Opening	Avg. Opening
2.20	2	2*-4.10	-4.10
2.10	3	3*-4.10	-4.10
2.00	6 **(11)**	6*-4.10	-4.10
1.90	9 **(20)**	8*-4.10, 1*-11.10	-4.88
1.80	18	15*-4.10, 1*-10.60, 2*-11.10	-5.24
1.70	35 **(73)**	26*-4.10, 1*-4.30, 1*-8.10, 1*-7.30, 2*-10.60, 4*-11.10	-5.25
1.60	55 **(128)**	38*-4.10, 2*-4.30, 2*-7.30, 2*-8.10, 3*-10.60, 8*-11.10	-5.74

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
11	2.06	-4.10
20	1.99	-4.45
73	1.80	-5.14

128	1.72	-5.40

U-Allele

Barrier	#LM	Opening	Avg. Opening
2.30	1	-11.10	-11.10
2.20	4	2*-11.10, 2*-7.30	-9.20
2.10	6 **(11)**	3*-11.10, 3*-7.30	-9.20
2.00	14 **(25)**	7*-11.10, 7*-7.30	-9.20
1.90	18	9*-11.10, 9*-7.30	-9.20
1.80	39 **(82)**	19*-11.10, 19*-7.30, -6.80	-9.14
1.70	66 **(148)**	32*-11.10, 32*-7.30, 2*-6.80,	-9.13

U-Allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
11	2.15	-9.37
25	51.7 2.07	-9.28
82	156.1 1.90	-9.19
148	1.81	-9.16

Ordered by Opening Energy

C-Allele

Opening	#LM	Barrier	Avg. Barrier
-3.70	14	1.10, 2*1.00, 3*0.90, 8*0.80	0.87
-3.80	28	6*1.00, 8*0.90, 14*0.80	0.87
-4.10	3,020	2*2.20, 3*2.10, 6*2.00, 8*1.90, 15*1.80, 26*1.70, 38*1.60, 53*1.50, 80*1.40, 104*1.30, 302*1.20, 1.14, 278*1.10, 2*1.04, 408*1.00, 3*0.94, 801*0.90, 7*0.84, 883*0.80	1.00

U-Allele

Opening	#LM	Barrier	Avg. Barrier
-2.90	1	0.90	0.90
-5.80	132	1.60, 2*1.50, 3*1.40, 6*1.30, 11*1.20, 18*1.10, 33*1.00, 58*0.90	1.02

11. HOXB5 (L = 952)

NM_002147.3 vs. rs9299
miRNA: miR-7
SNP Pos: 141
Target site: 126 - 154

RNAsubopt and Barrier Setting

Offset	Barrier	G-allele	A-allele	G-allele	A-allele
2.0	0.8	1,383,237	480,436	6,481	3,746

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
G-allele	-21.90	-0.49	-13.61	-8.29
A-allele	-22.40	-1.39	-11.20	-11.20

Local Minima

MFE	G-allele	A-allele
#Pairings	15	15
Opening Energy	-17.60	-18.30
Structure Energy	-302.70	-302.40
MFE Barrier	2.00	2.00
Minima		
Less Pairings	0.2% 16	1.4% 54
Equal Pairings	99.8% 6,464	98.6% 3,691
Greater Pairings	0	0
Identical to MFE	84.9% 5,504	98.6% 3,691
Minima (- MFE)	**6,480**	**3,745**
Avg. Opening of All Local Minima (excluding MFE)	**-17.65**	**-18.21**

#Pairings	#Minima G-allele	Target Site Approx. Avg.	#Minima A-allele	Target Site Approx. Avg.
9	-	-	1.44% 54	-12.20
12	0.2% 16	-13.95	-	-
15	**99.8% 6,464**	**-17.65**	**98.6% 3,691**	**-18.30**
Total Minima	6,480		3,745	
Avg. Barrier	1.00		0.94	

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	G-allele	A-allele
Less	4.0% 262	1.4% 54
Equal	84.9% 5,504	98.6% 3,691
Less or Equal	88.98% 5,766	100% 3,745

N+ Test

G-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-302.30	58 **(109)**	58*1.60	1.60	4*-18.30, 54*-17.60	-17.65
-302.40	22 **(51)**	22*1.70	1.70	4*-18.30, 18*-17.60	-17.73
-302.50	12 **(29)**	12*1.80	1.80	12*-17.60	-17.60
-302.60	12 **(17)**	12*1.90	1.90	12*-17.60	-17.60
-302.70	5	5*2.00	2.00	5*-17.60	-17.60
MFE: -302.70	1	2.00	2.00	-17.60	-17.60

G-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
17	-5144.7/17 -302.63	32.8/17 1.93	299.2/17 -17.60
29	-8774.7/29 -302.58	54.4/29 1.88	-510.4/29 -17.60
51	-15427.5/51 -302.50	91.8/51 1.80	-900.4/61 -17.65
109	-32960.9/109 -302.40	184.6/109 1.69	-1924/109 -17.65

A-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-301.90	36 **(111)**	32*1.50, 2*1.40, 2*1.30	1.48	36*-18.30	-18.30
-302.00	44 **(75)**	38*1.60, 4*1.40, 2*1.30	1.57	44*-18.30	-18.30
-302.10	12 **(31)**	6*1.70, 2*1.40, 4*0.80	1.35	12*-18.30	-18.30
-302.20	12 **(19)**	10*1.80, 2*1.40	1.73	12*-18.30	-18.30
-302.30	4	4*1.90	2.00	4*-18.30	-18.30
-302.40	3	3*2.00	2.00	3*-18.30	-18.30
MFE: -302.40	1	2.00	2.00	-18.30	-18.30

A-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
19	-5742.8/19 -302.25	34.4/19 1.81	347.7/19 -18.30
31	-9368/31	50.6/31	-567.3/31

	-302.19	1.63	-18.30
75	-22656/75 -302.08	119.6/75 1.59	-1372.5/75 -18.30
111	-33524.4/111 -302.02	173/111 1.56	-2031.3/111 -18.30

Ordered by Deepest Local Minima

G-allele

Barrier	#LM	Opening	Avg. Opening
2.00	5	5*-17.60	-17.60
1.90	12	12*-17.60	-17.60
1.80	12	12*-17.60	-17.60
1.70	22	22*-17.60	-17.60
1.60	54	54*-17.60	-17.60

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	19.5/10 1.95	-176/10 -17.60
20	38.2/20 1.91	-352/20 -17.60
60	106.2/60 1.77	-1056/60 -17.60
100	170.2/100 1.70	-1760/100 -17.60

A-allele

Barrier	#LM	Opening	Avg. Opening
2.00	3	3*-18.30	-18.30
1.90	4	4*-18.30	-18.30
1.80	10	10*-18.30	-18.30
1.70	6	6*-18.30	-18.30
1.60	38	38*-18.30	-18.30
1.50	32	32*-18.30	-18.30
1.40	114	114*-18.30	-18.30

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	19/10 1.90	-183/10 -18.30
20	36.7/20 1.84	-366/20 -18.30
60	101/60 1.68	-1098/60 -18.30
100	160.4/100 1.60	-1830/100 -18.30

Ordered by Opening Energy

G-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-13.95	16	2*1.10, 6*1.00, 8*0.90	0.96
-17.00	246	6*1.40, 12*1.30, 12*1.20, 18*1.10, 54*1.00, 102*0.90, 42*0.80	0.97

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	10/10 1.00	-139.5/10 -13.95
20	21/20 1.05	-291.2/20 -14.56
60	69.2/60 1.15	-971.2/60 -16.19
100	109.6/100 1.10	-1651.2/100 -16.51

A-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-12.20	54	6*1.10, 12*1.00, 12*0.90, 24*0.80	0.90
-18.30	3691	3*2.00, 4*1.90, 10*1.80, 6*1.70, 38*1.60, 32*1.50, 114*1.40, 96*1.30, 246*1.20, 222*1.10, 522*1.00, 680*0.90, 1718*0.80	0.94

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	10.6/10 1.06	-122/10 -12.20
20	20.4/20 1.02	-244/20 -12.20
60	60.3/60 1.01	-768.6/60 -12.81
100	127.2/100 1.27	-1500.6/100 -15.01

12. RAD51 (L = 978)

NM_002875.4 vs. rs7180135
miRNA: miR-197-3p
SNP Pos: 718
Target site: 707 - 725

RNAsubopt and Barrier Setting

Offset	Barrier	G-allele	A-allele	G-allele	A-allele
2.2	1.2	9,020,624	11,874,784	3,850	6,291

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
G-allele	-28.60	-0.56	-14.30	-14.31
A-allele	-22.00	-1.20	-15.00	-7.00

Local Minima

MFE	**G-allele**		**A-allele**	
#Pairings	6		6	
Opening Energy	-6.60		-4.70	
Structure Energy	-285.90		-284.00	
MFE Barrier	2.20		2.20	
Minima				
Less Pairings	0.0%	0	0.0%	0
Equal Pairings	100%	3,849	93.9%	5,907
Greater Pairings	0.0%	0	6.1%	383
Identical to MFE	100%	3,849	93.3%	5,867
Minima (- MFE)	**3,849**		**6,290**	
Avg. Opening of All Local Minima (excluding MFE)	**-6.60**		**-4.75**	

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	G-allele		A-allele	
Less	0.0%	0	0.6%	40
Equal	100%	3,849	93.3%	5,867
Less or Equal	**100%**	**3,849**	**93.9%**	**5,907**

N+ Test

G-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-285.40	204 **(376)**	203*1.70, 1.40	1.70	204*-6.60	-6.60
-285.50	134 **(172)**	99*1.80, 32*1.70, 3*1.20	1.76	134*-6.60	-6.60
-285.60	33 **(38)**	24*1.90, 8*1.70, 1.20	1.83	33*-6.60	-6.60

-285.70	2	2.00, 1.80	1.90	2*-6.60	-6.60
-285.90	3	2*2.10, 1.80	2.00	3*-6.60	-6.60
MFE: -285.90	1	2.20	2.20	-6.60	-6.60
-107618.1 Avg. -285.46	**377**	**655.10**	**1.74**	**-2488.19**	**-6.60**

G-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
10	-2857.1/10 -285.71	19.3/10 1.93	66/10 6.60
38	-10853.6/38 -285.62	70.2/38 1.85	-250.8/38 -6.60
172	-49110.6/172 -285.50	306.4/172 1.78	-1135.2/172 -6.60
376	-107332.2/376 -285.50	652.9/376 1.74	-2481.59/376 -6.60

A-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-283.50	207 **(379)**	204*1.70, 1.40, 2*1.20	1.69	207*-4.70	-4.70
-283.60	134 **(172)**	99*1.80, 32*1.70, 3*1.20	1.76	134*-4.70	-4.70
-283.70	33 **(38)**	24*1.90, 8*1.70, 1.20	1.83	33*-4.70	-4.70
-283.80	2	2.00, 1.80	1.90	2*-4.70	-4.70
-283.90	3	2*2.10, 1.80	2.00	3*-4.70	-4.70
MFE: -284.00	1	2.20	2.20	-4.70	-4.70
-107752.3 Avg. -283.60	**380**	**659.20**	**1.73**	**1786.89**	**-4.70**

A-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
10	-2857.1/10 -283.78	19.3/10 1.93	-47/10 -4.70
38	-10781.4/38 -283.70	70.2/38 1.85	-178.6/20 -4.70
172	-48783.79/172 -283.63	306.4/172 1.78	-808.4/172 -4.70
379	-107468.3/379 -283.60	657/379 1.73	-1782.19/379 -4.70

Ordered by Deepest Local Minima

G-allele

Barrier	#LM	Opening	Avg. Opening
2.10	2	2*-6.60	-6.60
2.00	1	-6.60	-6.60
1.90	24	24*-6.60	-6.60
1.80	101	101*-6.60	-6.60

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	19.5/10 1.95	-66/10 -6.60
20	38.5/20 1.93	-132/20 -6.60
60	91.5/60 1.53	-396/60 -6.60
100	183.2/100 1.83	-660/100 -6.60

A-allele

Barrier	#LM	Opening	Avg. Opening
2.10	2	2*-4.70	-4.70
2.00	1	-4.70	-4.70
1.90	24	24*-4.70	-4.70
1.80	101	101*-4.70	-4.70

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	19.5/10 1.95	-47/10 -4.70
20	38.5/20 1.93	-94/20 -4.70
60	91.5/60 1.53	-282/60 -4.70
100	183.2/100 1.83	-470/100 -4.70

Ordered by Opening Energy

G-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-6.60	3,849	2*2.10, 2.00, 24*1.90, 101*1.80, 243*1.70, 238*1.60, 446*1.50, 914*1.40, 1306*1.30, 574*1.20	1.39

G-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	19.5/10 1.95	-66/10 -6.60
20	46.1/20 2.31	-132/20 -6.60
60	111.2/60 1.85	-396/60 -6.60
100	183.2/100 1.83	-660/100 -6.60

A-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-3.70	1	1.20	1.20
-4.00	39	1.50, 3*1.40, 2*1.30, 33*1.20	1.23
-4.70	5867	2*2.10, 2.00, 24*1.90, 101*1.80, 243*1.70, 238*1.60, 446*1.50, 914*1.40, 1306*1.30, 2592*1.20	1.33

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	13.1/10 1.31	-39.7/10 -3.97
20	25.1/20 1.26	-79.7/20 -3.99
60	87.6/60 1.46	-253.7/60 -4.23
100	160.3/100 1.60	-441.7/100 -4.42

13. ORAI1 (L = 1034)

NM_032790.3 vs. rs76753792
miRNA: miR-519a-3p
SNP Pos: 86

	C-allele	U-allele
Target Site:	69 - 88	81 - 102

RNAsubopt and Barrier Setting

Offset	Barrier	C-allele	U-allele	C-allele	U-allele
2.0	0.9	223,532	559,335	577	1,332

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
U-allele	-17.90	-0.02	-23.42	5.62
C-allele	-16.70	-4.32	-9.49	-7.21

Local Minima

MFE	C-allele		U-allele	
#Pairings	6		16	
Opening Energy	-5.40		-27.21	
Structure Energy	-403.50		-405.20	
MFE Barrier	2.00		2.00	
Minima				
Less Pairings	0.0%	0	1.4%	18
Equal Pairings	100%	576	98.6%	1,313
Greater Pairings	0.0%	0	0.0%	0
Identical to MFE	100%	576	49.3%	656
Minima (- MFE)	**576**		**1,331**	
Avg. Opening of All Local Minima (excluding MFE)	**-5.40**		**-27.20**	

#Pairings	#Minima C-allele		Target Site Approx. Avg.	#Minima U-allele		Target Site Approx. Avg.
5	**100%**	**576**	**-5.40**	-		-
...	-		-	-		-
15	-		-	1.4%	18	-26.29
16	-		-	**98.65**	**1,313**	**-27.21**
Total Minima	**100%**	**576**	**-5.40**	**100%**	**1,331**	**-27.20**
Avg. Barrier	**1.16**			**1.12**		

Local Minima with Less Pairings

	C-allele	U-allele
#LM Less Pairings:	0.0%	1.4% 18
Avg. Target Site Energy	-	-26.29
Less Pairings Avg. Barrier Height	-	1.02

Local Minima with Less/Equal Approx. Target Site Energy than the MFE

	C-allele		U-allele	
Less	0.0%	0	1.4%	18
Equal	100%	576	98.6%	1,313
Less or Equal	**100%**	**576**	**100%**	**1,331**

C-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-402.90	35 **(132)**	21*1.40, 10*1.10, 4*1.00	1.27	35*-5.40	-5.40
-403.00	32	18*1.50, 4*1.10, 10*1.00	1.29	36*-5.40	-5.40
-403.10	26 **(65)**	22*1.60, 4*1.00	1.51	26*-5.40	-5.40
-403.20	22 **(39)**	12*1.70, 10*0.90	1.34	22*-5.40	-5.40
-403.30	4	4*1.80	1.80	4*-5.40	-5.40
-403.40	10 **(13)**	10*1.90	1.90	10*-5.40	-5.40
-403.50	3	3*2.00	2.00	3*-5.40	-5.40
MFE: -403.50	1	2.00	2.00	-5.40	-5.40
-53609.7 Avg. -403.10	**133**	**188.6**	**1.42**	**-718.20**	**-5.40**

C-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
39	-15728.1/39 -403.30	61.6/39 1.58	-210.6/20 -5.40
65	-26208.7/65 -403.20	100.8/65 1.55	-351/65 -5.40
132	-53206.2/132 -403.10	186.6/132 1.41	-712.8/100 -5.40

U-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-404.80	52 **(111)**	44*1.60, 8*1.00	1.51	52*-27.21	-27.21
-404.90	24 **(59)**	24*1.70	1.70	24*-27.21	-27.21
-405.00	8	8*1.80	1.80	8*-27.21	-27.21
-405.10	20 **(27)**	20*1.90	1.90	20*-27.21	-27.21
-405.20	7	7*2.00	2.00	7*-27.21	-27.21
MFE: -405.20	1	2.00	2.00	-27.21	-27.21
-45350.8 Avg. -404.92	**112**	**187.6**	**1.68**	**-3047.52**	**-27.21**

U-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
111	-44945.6/111 -404.90	185.6/111 1.67	-3020.31/111 -27.21
59	-23896/59 -405.00	107.2/59 1.82	-1605.39/59 -27.21
27	-10938.4/27 -405.10	52/27 1.93	-737.67/20 -27.21

Ordered by Deepest Local Minima

C-allele

Barrier	#LM	Opening	Avg. Opening
2.00	3	3*-5.40	-5.40
1.90	10	10*-5.40	-5.40
1.80	4	4*-5.40	-5.40
1.70	12	12*-5.40	-5.40
1.60	22	22*-5.40	-5.40
1.50	18	18*-5.40	-5.40
1.40	21	21*-5.40	-5.40
1.30	38	38*-5.40	-5.40

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
13	25/13 1.92	-70.2/13 -5.40
29	52.6/29 1.81	-156.6/29 -5.40
69	114.8/69 1.66	-372.6/69 -5.40
128	193.6/128	-691.19/100

	1.51	-5.40

U-allele

Barrier	#LM	Opening	Avg. Opening
2.00	7	7*-27.21	-27.21
1.90	20	20*-27.21	-27.21
1.80	8	8*-27.21	-27.21
1.70	24	24*-27.21	-27.21
1.60	44	44*-27.21	-27.21

U-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
7	14/7 2.00	-190.47/10 -27.21
27	52/27 1.93	-734.67/27 -27.21
59	107.2/59 1.82	-1632.6/59 -27.21
103	177.6/100 1.72	-2802.63/100 -27.21

Ordered by Opening Energy

C-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-5.40	576	3*2.00, 10*1.90, 4*1.80, 12*1.70, 22*1.60, 18*1.50, 21*1.40, 38*1.30, 4*1.20, 131*1.10, 181*1.00, 83*1.20, 49*0.90	1.16

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	19.3/10 1.93	-54/10 -5.40
20	37.3/20 1.87	-108/20 -5.40
60	101.3/60 1.69	-324/60 -5.40
100	157.2/100 1.57	-540/100 -5.40

U-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-26.31	18	4*1.10, 14*1.00	1.02
-27.21	1313	7*2.00, 20*1.90, 8*1.80, 24*1.70, 44*1.60, 36*1.50, 42*1.40, 76*1.30, 174*1.20, 242*1.10, 382*1.00, 258*0.90	1.13

U-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	10.4/10 1.04	-263.1/10 -26.31
20	22.4/20 1.12	-528/20 -26.40
60	96.7/60 1.61	-1616.4/60 -26.94
100	162.4/100 1.62	-2704.8/100 -27.05

14. RAP1 (L = 1078)

NM_02884.2 vs. rs6573
miRNA: miR-196a
SNP Pos: 366
Target Site: 348 - 370

RNAsubopt and Barrier Setting

Offset	Barrier	C-allele	A-allele	C-allele	A-allele
2.0	0.9	1,689,428	1,689,428	238	238

STarMir Target Site Energies

	dG_hybrid	dG_nucl	dG_open	dG_total
C-allele	-16.70	-3.97	-5.40	-11.30
A-allele	-21.30	-6.97	-5.05	-16.25

Local Minima

MFE	**C-allele**		**A-allele**	
#Pairings	12		12	
Opening Energy	-5.10		-5.10	
Structure Energy	-197.80		-197.80	
MFE Barrier	2.00		2.00	
Minima				
Less Pairings	0.0%		0.0%	
Equal Pairings	100%	237	100%	237
Greater Pairings	0.0%		0.0%	
Identical to MFE	100%	237	100%	237
Minima (- MFE)	**237**		**237**	
Avg. Opening of All Local Minima (excluding MFE)	**-5.10**		**-5.10**	

C-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-196.90	46 **(143)**	46*1.10	1.10	46*-5.10	-5.10
-197.00	30	30*1.20	1.20	30*-5.10	-5.10
-197.10	28 **(67)**	22*1.30, 6*0.90	1.21	28*-5.10	-5.10
-197.20	12	10*1.40, 2*1.30	1.38	12*-5.10	-5.10
-197.30	10 **(27)**	10*1.50	1.50	10*-5.10	-5.10
-197.40	8	6*1.60, 2*1.50	1.58	8*-5.10	-5.10
-197.50	6	6*1.70	1.70	6*-5.10	-5.10
-197.70	2	2*1.90	1.90	2*-5.10	-5.10
-197.80	1	2.00	2.00	-5.10	-5.10
MFE: -197.80	1	2.00	2.00	-5.10	-5.10

-28380.8 Avg. -197.09	**144**	**182.8**	**1.27**	**-734.4**	**-5.10**

C-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
143	-28183/143 -197.08	180.8/143 1.26	-729.3/143 -5.10
67	-13215.6/67 -197.25	94.2/67 1.41	-341.7/67 -5.10
27	-5330.4/27 -197.42	43.6/27 1.61	-137.7/27 -5.10

A-allele - Local Minima Distribution

E	#LM	Barrier	Avg. Barrier	Opening	Avg. Opening
-196.90	46 **(143)**	46*1.10	1.10	46*-5.10	-5.10
-197.00	30	30*1.20	1.20	30*-5.10	-5.10
-197.10	28 **(67)**	22*1.30, 6*0.90	1.21	28*-5.10	-5.10
-197.20	12	10*1.40, 2*1.30	1.38	12*-5.10	-5.10
-197.30	10 **(27)**	10*1.50	1.50	10*-5.10	-5.10
-197.40	8	6*1.60, 2*1.50	1.58	8*-5.10	-5.10
-197.50	6	6*1.70	1.70	6*-5.10	-5.10
-197.70	2	2*1.90	1.90	2*-5.10	-5.10
-197.80	1	2.00	2.00	-5.10	-5.10
MFE: -197.80	1	2.00	2.00	-5.10	-5.10
-28380.8 Avg. -197.09	**144**	**182.8**	**1.27**	**-734.4**	**-5.10**

A-allele - N+ Averages

Test #LM-MFE	Structure Energy	Barrier	Opening Energy
27	-5330.4/27 -197.42	43.6/27 1.61	-137.7/27 -5.10
67	-13215.6/67 -197.25	94.2/67 1.41	-341.7/67 -5.10
143	-28183/143 -197.08	180.8/143 1.26	-729.3/143 -5.10

Ordered by Deepest Local Minima

C-allele

Barrier	#LM	Opening	Avg. Opening
2.00	1	-5.10	-5.10

1.90	2	2*-5.10	-5.10
1.70	6	6*-5.10	-5.10
1.60	6 **(15)**	6*-5.10	-5.10
1.50	12 **(27)**	12*-5.10	-5.10
1.40	10	10*-5.10	-5.10
1.30	24 **(61)**	24*-5.10	-5.10
1.20	30	30*-5.10	-5.10
1.10	46 **(137)**	46*-5.10	-5.10

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
15	25.6/15 1.71	-76.5/15 -5.10
27	43.6/27 1.61	-137.7/27 -5.10
61	88.1/61 1.46	-311.1/61 -5.10
137	175.4/128 1.28	-698.7/137 -5.10

A-allele

Barrier	#LM	Opening	Avg. Opening
2.00	1	-5.10	-5.10
1.90	2	2*-5.10	-5.10
1.70	6	6*-5.10	-5.10
1.60	6	6*-5.10	-5.10
1.50	12	12*-5.10	-5.10
1.40	10	10*-5.10	-5.10
1.30	24	24*-5.10	-5.10
1.20	30	30*-5.10	-5.10
1.10	46	46*-5.10	-5.10

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
15	25.6/15 1.71	-76.5/15 -5.10
27	43.6/27 1.61	-137.7/27 -5.10
61	88.1/61 1.46	-311.1/61 -5.10
137	175.4/128 1.28	-698.7/137 -5.10

Ordered by Opening Energy

C-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-5.10	237	2.00, 2*1.90, 6*1.70, 6*1.60, 12*1.50, 10*1.40, 24*1.30, 30*1.20, 46*1.10, 77*1.00, 22*0.90	1.45

C-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	17.6/10 1.76	-51/10 -5.10
20	33.1/20 1.66	-102/20 -5.10
60	87.5/60 1.46	-306/60 -5.10
100	134.7/100 1.35	-510/100 -5.10

A-allele - Opening Energy

Opening	#LM	Barrier	Avg. Barrier
-5.10	237	2.00, 2*1.90, 6*1.70, 6*1.60, 12*1.50, 10*1.40, 24*1.30, 30*1.20, 46*1.10, 77*1.00, 22*0.90	1.45

A-allele - Averages

Test #LM-MFE	Avg. Barrier	Avg. Opening Energy
10	17.6/10 1.76	-51/10 -5.10
20	33.1/20 1.66	-102/20 -5.10
60	87.5/60 1.46	-306/60 -5.10
100	134.7/100 1.35	-510/100 -5.10

Appendix D

RNA-binding Proteins and microRNAs

RBP-miRNA Cooperation

MYC

3' UTR Length: 476

- miRNA let-7 requires HuR to reduce expression of MYC.
- The HuR and miRNA binding site are >30 nt apart.
- Two overlapping HuR binding sites.
  CUUACCAUCUUUUUUUUUUCUUUAACAG
  UAACAGAUUUGUAUUUAAGAAUUGUUUUU

Lafon et al. Developmental expression of AUF1 and HuR, two c-myc mRNA binding proteins. Oncogene. 1998 2;16(26):3413-21.

Kim HH, Kuwano Y, Srikantan S, Lee EK, Martindale JL, Gorospe M. HuR recruits let-7/RISC to repress c-Myc expression. Genes Dev. 2009 Aug 1; 23(15):1743-8.

Menachem J Gunzburg et al. Cooperative interplay of let-7 mimic and HuR with MYC. RNA Cell Cycle, 14:17. 2015.

CDKN1B/P27

3' UTR Length: 1,344

- Suppression of P27 by miR-221 requires PUM1. "PUM1 knockdown abolished miR-221 function, but loss of its binding sites on the p27-3' UTR did not."
- PUM1 binding changes accessibility of miR-221 site.
- P27 3' UTR contains two evolutionary conserved PUM recognition elements, one located nearby to miR-221 and 222 target sites.
- Also experimental evidence, RBP DND1 binding reduces miR-221 by reducing accessibility.

Kedde, M et al. A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. Nat. Cell Biol. 2010, 12, 1014–1020.

Kedde et al. RNA-Binding Protein Dnd1 Inhibits MicroRNA Access to Target mRNA. Cell, 2007, 131, 1273–1286

RhoB

3' UTR Length: 1,389

- In vivo evidence that HuR required for miR-19 suppression.
- HuR site position 823: AUUUAUUUA
- miR-19 site position 841.

Glorian V, Maillot G, Polès S, Iacovoni JS, Favre G, Vagner S. HuR-dependent loading of miRNA RISC to the mRNA encoding the Ras-related small GTPase RhoB controls its translation during UV-induced apoptosis. Cell Death Differ. 2011; 18:1692–1701

GNPDA1

3’ UTR Length: 1,360

- PTB promotes miRNA activity on gene GNPDA1 by change of secondary structure.

Xue, Y.; Ouyang, K.; Huang, J.; Zhou, Y.; Ouyang, H.; Li, H.; Wang, G.; Wu, Q.; Wei, C.; Bi, Y.; et al. Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. Cell 2013, 152, 82–96.

E2F3

3’ UTR Length: L = 3,285

- PUM promotes suppression of E2E3 by miR-503.

Miles, W.O.; Tschop, K.; Herr, A.; Ji, J.Y.; Dyson, N.J. Pumilio facilitates miRNA regulation of the E2F3 oncogene. Genes Dev. 2012, 26, 356–368.

RBP-miRNA Competition

TP53

- Binding of HuR prevents miR-125b repression of P53.

Deepika Ahuja et al. Interplay between RNA-binding protein HuR and microRNA-125b regulates p53 mRNA translation in response to genotoxic stress. RNA Bio. 2016, Vol. 13, NO. 11, 1152–1165.

COX-2

- HuR binding to 3’ UTR prevents miR-16 activity.
- HuR and miRNA sites reported to be nearby.

Young L, et al. The mRNA Stability Factor HuR Inhibits MicroRNA-16 Targeting of Cyclooxygenase-2. Mol Cancer Res. 2012 Jan; 10(1): 167–180.

PDCD4

- HuR binds to 3’ UTR preventing miR-21 repression of PDCD4.

D K Poria et al. RNA-binding protein HuR sequesters microRNA-21 to prevent translation repression of proinflammatory tumor suppressor gene programmed cell death 4. Oncogene (2016) 35, 1703–1715.

CAT-1

- Binding of HuR releases miR-122 binding.

Bhattacharyya SN, Habermacher R, Martine U, Closs EI, Filipowicz W. Relief of microRNA-mediated translational repression in human cells subjected to stress. Cell 2006; 125:1111